AFRL-IF-RS-TR-2000-144
Final Technical Report
November 2000

ULTRA-HIGH-CAPACITY ROUTER SWITCHING FABRICS USING PHOTONIC TECHNOLOGIES

Drexel University

Sponsored by
Defense Advanced Research Projects Agency
DARPA Order No. H109

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

AIR FORCE RESEARCH LABORATORY
INFORMATION DIRECTORATE
ROME RESEARCH SITE
ROME, NEW YORK

This report has been reviewed by the Air Force Research Laboratory, Information Directorate, Public Affairs Office (IFOIPA) and is releasable to the National Technical Information Service (NTIS). At NTIS it will be releasable to the general public, including foreign nations.

AFRL-IF-RS-TR-2000-144 has been reviewed and is approved for publication.

APPROVED:

PRISCILLA A. CASSIDY
Project Engineer

FOR THE DIRECTOR:

WARREN H. DEBANY, Technical Advisor
Information Grid Division
Information Directorate

If your address has changed or if you wish to be removed from the Air Force Research Laboratory Rome Research Site mailing list, or if the addressee is no longer employed by your organization, please notify AFRL/IFGA, 525 Brooks Road, Rome, NY 13441-4505. This will assist us in maintaining a current mailing list.

Do not return copies of this report unless contractual obligations or notices on a specific document require that it be returned.

ULTRA-HIGH-CAPACITY ROUTER SWITCHING FABRICS USING PHOTONIC TECHNOLOGIES

Stewart D. Personick,
Hongyuan Shi, and
Dinesh Shankar

Contractor: Drexel University
Contract Number: F30602-99-1-0012
Effective Date of Contract: 16 December 1998
Contract Expiration Date: 15 March 1999
Short Title of Work: Ultra-High-Capacity Router Switching Fabrics Using Photonic Technologies
Period of Work Covered: Dec 98 - Mar 99

Principal Investigator: Stewart D. Personick
Phone: (212) 895-1695
AFRL Project Engineer: Priscilla A. Cassidy
Phone: (315) 330-1887

This research was supported by the Defense Advanced Research Projects Agency of the Department of Defense and was monitored by Priscilla A. Cassidy, AFRL/IFGA, 525 Brooks Road, Rome, NY.

REPORT DOCUMENTATION PAGE			Form Approved OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	NOVEMBER 2000	Final Dec 98 - Mar 99

4. TITLE AND SUBTITLE
ULTRA-HIGH-CAPACITY ROUTER SWITCHING FABRICS USING PHOTONIC TECHNOLOGIES

5. FUNDING NUMBERS
C - F30602-99-1-0012
PE - 62110E
PR - H109
TA - 00
WU - 01

6. AUTHOR(S)
Stewart D. Personick, Hongyuan Shi, and Dinesh Shankar

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Drexel University
Department of Electrical & Computer Engineering
3141 Chestnut Street
Philadelphia PA 19104

8. PERFORMING ORGANIZATION REPORT NUMBER
AFRL-IF-RS-TR-2000-144

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Defense Advanced Research Projects Agency
3701 North Fairfax Drive
Arlington VA 22203-1714

Air Force Research Laboratory/IFGA
525 Brooks Road
Rome NY 13441-4505

10. SPONSORING/MONITORING AGENCY REPORT NUMBER
AFRL-IF-RS-TR-2000-144

11. SUPPLEMENTARY NOTES
Air Force Research Laboratory Project Engineer: Priscilla A. Cassidy/IFGA/(315) 330-1887

12a. DISTRIBUTION AVAILABILITY STATEMENT
APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*
In this report we present an architecture and a design study of the feasibility of implementing an ultra-high-capacity router switching fabric for Next Generation Internets utilizing mostly photonic components.

14. SUBJECT TERMS
Photonic, Switch, Internet

15. NUMBER OF PAGES
60

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UL

Standard Form 298 (Rev. 2-89) (EG)
Prescribed by ANSI Std. 239.18
Designed using Perform Pro, WHS/DIOR, Oct 94

Table of Contents

Ultra-High-Capacity Router Switching Fabrics Using Photonic Technologies.....1

Executive Summary.....1

1.0 Introduction.....2

2.0 Requirements for an Ultra-High-Capacity Router Switching Fabric Using Photonic Technologies.....3

2.1 Layer 2 Network Management Functions.....3
2.2 IP Functions.....4
2.2.1 Basic Requirements and considerations.....5

3.0 Background on Photonic Switching Fabrics.....5

4.0 Design.....8
4.1 Packet Structure Design Tradeoffs.....8
4.2 Separation of Network Layer Router Address and Management (NLRAM) Information from the Incoming Packets.....9
4.3 Regeneration to Remove the Effects of Transmission of Packets from their Sources...11
4.4 Aligning Arriving Packets with the Router Master Clock.....11
4.5 Packets Which Are Competing for the Same Output Port.....13
4.6 The Main Switching Matrix.....17
4.7 Asynchronous Regeneration of Packets.....18
4.8 Adding NLRAM Information to the Outgoing Packets.....19
4.9 Putting It All Together.....19
4.10 Performance of the Switching Fabric.....19
4.11 Observations.....20

5.0 Second Generation Design: A Wavelength-Space-Wavelength Swithc.....20

6.0 Networking Issues.....21

7.0 Conclusions.....22

List of Figures

Figure	1	OC-3	23
Figure	2	Separate mechanisms for Placing NLRAM and payload information on the Optical layer (violates traditional protocol stack paradigm)	24
Figure	3a	Inserting NLRAM Information: WDM	25
Figure	3b	Inserting NLRAM Information: Overly modulation	26
Figure	3c	Inserting NLRAM Information: Repeated bits	27
Figure	4	Guard band between optical packets	28
Figure	5a	Separating NLRAM: WDM	29
Figure	5b	Separating NLRAM: Overlay modulation	30
Figure	5c	Separating NLRAM: Embedded in packet stream	31
Figure	5d	Separating NLRAM: Embedded in packet stream	32
Figure	6	7-Sage Delay Equalizer	33
Figure	7a	5 stage buffer: 0-31T	34
Figure	7b	N stage buffer 0-31 T	35
Figure	7c	1 X 8 Banyon Switch	36
Figure	7d	Buffer subsystem	37
Figure	8	8 X 8 Benes Matrix	38
Figure	9	Asynchronous Regenerator	39
Figure	10a	Adding NLRAM: WDM	40
Figure	10b	Adding NLRAM: Overlay modulation	41
Figure	10c	Adding NLRAM: Bit Stream Insertion	42
Figure	11	Overall Architecture	43
Figure	12	Performance (U=utilization)	44
Figure	13	Wavelength-Space-Wavelength Switching Matrix	45

Deliverable Version 1.1
April 14, 1999

Ultra-High-Capacity Router Switching Fabrics Using Photonic Technologies

This architectural/design study was performed under Air Force Research Laboratory Agreement No. **F30602-99-1-0012**

Principal Investigator: Stewart D. Personick, Drexel University
Contributing Investigators: Hongyuan Shi, Drexel University; Dinesh Shankar, Drexel University

Executive Summary

In this report, we present an architecture and a design study of the feasibility of implementing an ultra high capacity router switching fabric for next generation Internets: utilizing (mostly) photonic components.

Our basic design concept (which includes several variations) is shown in figure 11.

Packets arriving at a router input port first pass through a WDM demultiplexer; in which network-layer router address and management (NLRAM) information, e.g., packet addresses, is stripped off. The remainder of each packet then enters a traditional regenerator, which removes the effects of transmission from the distant source.

Regenerated packets pass through a packet aligner, which aligns the packet stream with the local packet clock. The aligner also includes an optical amplifier to compensate for the insertion loss of the aligner.

Packets then pass through a buffer subassembly consisting of: a 1 x 32 switch; 32 available, 0-31 packet delay buffers; and a 32 x 1 switch. Depending upon the design of the 0-31 packet delay buffers, each of these buffers may incorporate an optical amplifier. The output of the buffer subassembly includes an optical amplifier.

Packets then enter the 128 x 128 main switching matrix (a rearrangably non-blocking Benes matrix). The main switching matrix may require optical amplifiers between some subassemblies of the main switching matrix to compensate for cumulative insertion losses.

The outputs of the main switching matrix direct packets to a set of 128 asynchronous regenerators. After each regenerator, NLRAM information is added using a wavelength multiplexer.

We conclude that, using existing technologies, it is feasible to implement an ultra high capacity (1 billion packets per second) router switching fabric: utilizing (mostly) photonic components. Such a router holds out the promise of significantly reduced physical design challenges, compared to traditional electronic router switching fabrics.

We also conclude that advances in optical networking technologies (wavelength translators and wavelength selective crosspoints) would enable the implementation of a wavelength-space-wavelength (W-S-W) switching fabric. A W-S-W switching fabric would require a substantially reduced number of optical components and interconnections. We see no major obstacles, related to implementing network management functionality, which would be introduced by the proposed use of long, fixed length packets, and the absence of a SONET layer.

This architectural/design study was performed under Air Force Research Laboratory Agreement No. F30602-99-1-0012

1.0 Introduction

The ongoing rapid expansion of the Internet is leading to requirements for networking technologies that can support aggregate throughputs of greater than 4 Terabits per second (~ 1 billion packets per second, each containing an average of 512 bytes). High speed electronic circuits, required to create switching fabrics for ultra-high-capacity routers, are also advancing rapidly in their capabilities; but are constrained by physical design issues, as circuit speeds and circuit densities increase, related to heat generation and difficulties in implementing point-to-point interconnections. For more than two decades, researchers have been investigating approaches for creating ultra-high-capacity switching fabrics using photonic technologies and architectures. The challenge of using photonic technologies is two-fold. First, one requires photonic components with the appropriate properties. For example, switching components with low insertion losses and adequate ratios of extinction into unwanted paths. Second, one needs a system architecture for the switching fabric, and a network architecture for the overall network, which can capitalize on the favorable properties of photonic technologies, while avoiding problems associated with the unfavorable properties of photonic components. For example, some photonic components, such as fibers and some types of photonic switches, can accommodate optical signals which have been modulated at very high data rates; independent of how those signals have been modulated ("optical transparency"). Optical fibers can be used to create delay lines with very large delay - data rate products. Optical signals travelling in adjacent optical fibers do not interfere with each other. On the other hand, photonic components are generally far inferior to electronic components for performing complex digital computation operations and for high-speed read-write random access storage of individual bits of information.

In this report, we will provide an overview of the requirements that a successful photonic switching fabric architecture/design must satisfy; and we will provide the results of an architectural/design study that was performed under Air Force Research Laboratory Agreement No. F30602-99-1-0012

We will conclude that an ultra-high-capacity photonic switching fabric for use in next generation Internets is feasible; without using wavelength selective switching components. We will conclude that a 2nd generation switching fabric, employing wavelength selective switching components and wavelength translation components (to create a wavelength-space-wavelength switching fabric architecture) would produce a substantial reduction in the number of individual components required to realize the switching fabric. This would result in an associated potential reduction: in the physical size of the fabric; the cumulative loss, statistical loss variability, and crosstalk in the associated optical components; and in the required number of optical interconnections between components.

2.0 Requirements for an Ultra-High-Capacity Router Switching Fabric Using Photonic Technologies

2.1 Layer 2 Network Management Functions

In order to understand the requirements that a (largely) photonic router switching fabric should/must meet, it is helpful to do a brief review of current generation IP networks; focusing only on those aspects which appear relevant to this report. Refer for figure 1.

Figure 1 is a representative example of how an IP network might be implemented with today's technologies and network architecture. We have placed a current-generation, high-speed, high-capacity backbone router at the center of figure 1. This backbone router is connected to other backbone routers on the backbone network, and to "regional" routers associated with networks that are interconnected using the backbone network. Edge routers are associated with the local area networks and campus networks of customers who subscribe to the transport services of the regional or backbone networks.

The physical links connecting other routers to the highlighted backbone router in figure 1 are shown as SONET OC-3 (155 Mbps), SONET OC-12 (622 Mbps), and SONET OC-48 (2.5 Gbps) transmission systems. The reader should keep in mind that figure 1 represents a typical present-generation IP backbone network, and that future generation networks may or may not utilize SONET as a layer 1-2 protocol. For simplicity, in this discussion, we shall assume that IP (Internet Protocol) packets are directly inserted into the SONET payload, without the use of ATM (asynchronous transfer mode) as an intermediate layer 2 protocol. The use of ATM as a layer 2 protocol between IP and SONET has no direct impact on the discussions that follow in this report; but is an important topic in other contexts regarding network architecture.

One role that is served by the SONET layer (among others) is to facilitate the conveyance of network management information between network "elements" (e.g., routers); which have been designed to: monitor/act on/insert this network management information into the SONET network-management-overhead information fields that accompany the payload. This network management information is used, for example, to coordinate recovery from network equipment or link failures. Any proposed next-generation networking approach must include a proposed mechanism for enabling network elements

(e.g., routers) to convey network management information that is currently conveyed in the SONET network management overhead fields.

Another network function that is enabled by the SONET network management overhead, and highlighted in figure 1, is the establishment of end-to-end "paths". Paths are composed of portions of the transmission capacities of transmission facilities that lie (in a path) between two specified end points. They produce an identified and managed end-to-end transmission capability between these two end points. Paths typically traverse multiple SONET (or other) transmission facilities. Network managers need to establish, monitor, and restore these end-to-end paths to meet customer needs; separate and apart from being able to monitor the individual transmission facilities/links that are utilized by these end-to-end paths. Any proposed next-generation networking approach should include a proposed mechanism for enabling the establishment of end-to-end paths made up of designated portions of concatenated transmission facilities. Such paths can be thought of as end-to-end permanent virtual circuits with specified quality-of-service requirements, which are monitored and maintained by network management systems.

2.2 IP Functions

The Internet Protocol (IP) has been remarkably successful because it was designed to be compatible with a wide variety of transmission media, and with systems which might be called upon to transport IP packets. The IP protocol is currently being updated in several ways:

- to accommodate a larger number of potential IP addresses (IPv6);
- to accommodate nomadic users ("mobile IP");
- to provide for encryption of point-to-point links (strictly speaking, not an IP-layer function in the networking protocol stack, but referred to as IPSec);
- to provide a mechanism that can support quality-of-service (QOS) guarantees across an Internet ("RSVP").

Some of these protocol changes/upgrades are well along the standardization process; but the marketplace will determine which of these become widely deployed. In some cases, standardization not withstanding, these protocol changes and upgrades are still being examined and debated, in terms of their potential benefits and the difficulties their introductions might imply. For example, the currently proposed methods for encrypting links in IP networks may cause problems with respect to access to network management information contained in the bit streams that flow through these proposed encrypted links. As we shall describe below, for high speed photonic switching fabrics to produce their potential benefits, it would be helpful if information required to route packets and information needed for network management purposes were readily accessible; i.e., without having to decrypt each incoming packet at its underlying bit rate. Thus, as the Internet moves toward its next generation, the tradeoffs, interactions, and compromises that need to be made, between the evolving IP protocol and the systems that will implement the protocol, are becoming more complex.

2.2.1 Basic requirements and considerations

The most basic requirement of the layer 3 IP protocol is to provide mechanisms to convey:

- the destination address of a packet, and, in the future, abbreviated addresses ("tags") associated with multi-packet "flows";
- priority information and/or other information related to a packet's quality of service requirements;
- other information that each router will have to access in order to make routing decisions.

Collectively, this information, which must be readily accessible at each router, corresponds to a number of bytes of information which we will refer to here as "**Network Layer Address and Router Management Information" (NLARM information).** This terminology is created here as a convenience for reference in this report, and has no other significance. A switching fabric and networking architecture utilizing photonic components much provide a mechanism for making NLARM information accessible at each router.

In IP networks, as they are implemented today, the NLARM information is represented by a set of bytes within the aggregate sequence of bytes that comprises a layer 3 packet. Access to the NLARM information is obtained by examining the packet on a bit-by-bit basis. In what follows, we are going to propose that the NLARM information should be associated with a packet in such a way as to make it accessible without examining the packet on a bit-by-bit basis. This is not a new idea, but a proposal to move in this direction is a proposal to depart from the well-established practice that is followed in implementing IP networks.

In particular, an important implication of this departure is that an IP packet (in which the NLARM information is segregated from the rest of the bytes in the packet) requires a split layer 1-2 networking facility which can accept (from layer 3) and transport these segregated portions of the packet. This segregation process causes a coupling of the implementations of protocol layers 1-3 which violates the "separation of concerns" among those layers which is traditionally a part of the networking protocol stack. Saying this another way, layers 1&2 now have to be designed specifically to accept two segregated sequences of bytes, rather than a single sequence of bytes. See figure 2.

3.0 Background on Photonic Switching Fabrics

Since the very successful initial deployment of fiber optic systems, in the early 1980's, there has been considerable interest in, enthusiasm for, and speculation regarding the feasibility and potential of photonic switching. Photonic switching *devices* have been demonstrated using a variety of designs and underlying physical phenomena. These include:

-mechanical switching devices; in which fibers, or components placed in the paths between fibers, are moved to cause the switching action

-electro-optic switching devices; in which a voltage applied to selected portions of the device creates fields within the device that change the optical properties of selected materials within the device. This electro-optic effect, results in the desired switching action. For example, one might modulate the mismatch between the phase velocities of two coupled optical waveguides formed on the surface of a switching device

-acousto-optic devices; in which one or more acoustic (pressure) waves is launched into the device by applying a periodically varying voltage across a set of electrodes (transducer) formed on the surface of the device. These acoustic waves, traveling through the device, create one or more optical gratings within the device. Turning the acoustically induced gratings on and off, or changing their pitches (periods) or directions causes the desired switching action when an optical beam passing through the gratings is diffracted

-all-optical devices; in which the application of an optical control signal to a switching device causes a change in the properties of a non-linear optical material within the device... thereby leading to switching action for another optical signal passing through the same device

-switched semiconductor amplifier arrays; in which paths through amplifying devices are turned on and off by applying currents to modulate the amplifiers on and off

-wavelength selective components, combined with tunable lasers; in which the selection of the wavelength of a tunable laser changes the path of the associated optical beam through a wavelength selective component (e.g., a grating)

With the exception of the all-optical switching devices, each of the above switching devices requires an electrical control signal (a voltage or a current) to set and/or change the state (i.e., the path toward which light is directed) of the switch. Even the all-optical devices typically have electrical control signals to turn optical control signals on and off.

Thus the use of photonic switching devices does not, in and of itself, eliminate the need for an electronic switching control subsystem that produces independent control signals for each switching element, which run at the full photonic device switching speeds. The collection of these control signals and the associated electronic circuits can be considered as a shadow switching fabric, in parallel with the optical switching fabric.

Photonic switching devices have critical dimensions, where interactions between optical signals and/or between optical signals and the device occur, that are required to be longer than the wavelength of the light signal being switched; and often hundreds or thousands of optical wavelengths (microns) long. The relevant dimensions of electronic switching devices are typically sub-micron size. As a result *, the control electrodes of photonic switching devices have capacitances that are typically larger than the capacitances of state-of-the-art electronic switching devices. This means that the quantity of charge that

must be added to, or removed from a photonic switching device to change its state is relatively large compared to that in a typical electronic device. As a result, the power dissipated in circuits and associated transmission lines that drive optical switching devices at a given switching rate is typically larger than the power dissipated in electronic switching devices running at the same switching rate (state changes per second).

*This is a somewhat qualitative argument, because the actual value of the input capacitance associated with a pair of control electrodes depends on several factors, including: the dielectric constant of the material between the electrodes and the spacing between the electrodes; not just the length of the electrodes.

While photonic switching devices can typically transmit optical signals with arbitrary modulation imposed upon them, the above explains why it is important, in practical switching fabric designs, to avoid rapid and frequent changes the states of optical switching devices that are part of such a fabric.

While electronic switching devices are typically non-linear in their responses to control signals (i.e., the exact value of the control signal is not critical provided that it is either above or below the device's threshold for switching), several types of photonic switching devices typically require relatively precise control voltages to set them in the desired state. Exceptions to this are semi-conductor optical amplifier switches and most mechanical photonic switches.

The precise voltage required to set the state in many photonic switch designs represents a practical concern with regard to the feasibility of a photonic switching fabric containing a large number of such switches.

Photonic switching devices introduce both nominal and statistical losses as signals pass through them. Even semiconductor amplifier switches introduce statistically variable gain to signals that are transmitted. Nominal and statistical losses are associated with the devices themselves, as well as coupling losses from fiber or free space interconnections into and out of the devices. Nominal losses may range from a fraction of a decibel to several decibels per switching device. Statistical losses may also range from a fraction of a decibel to several decibels. The specific values of nominal and statistical losses depend upon the device design and the number of inputs and outputs associated with a single device. Nominal and statistical losses through a switching fabric can accumulate to the point where the signal-to-noise ratio associated with an optical signal is too low for reliable recovery of the underlying information. Optical amplifiers or regenerative repeaters are required to prevent such situations from occurring. In addition, the accumulation of nominal and statistical losses can result in significant power level differences among signals travelling through the same switching fabric. Power level differences increase the vulnerabilities of the weaker signals to crosstalk from the stronger signals, if both weaker and stronger signals pass through a common switching device. Thus optical amplifiers may be required to reduce the power level differences among signals in the same switching fabric.

4.0 Design

In the sections below we shall provide the results of the switching fabric architectural design which was performed under this agreement. Our articulation of these results will begin with a review of the objectives and assumptions that we began with; and then proceed as follows:

-packet structure design and tradeoffs
-separation of Network Layer Router Address and Management (NLRAM) information from the incoming packets
-regeneration to remove the effects of transmission of packets from their sources
-alignment of arriving packets with the router master packet clock
-buffering of packets which are competing for the same output port
-the main packet switching fabric
-asynchronous regeneration of packets which have traversed the switching fabric
-adding NLARM information to the outgoing packets

4.1 Packet Structure Design Tradeoffs

Following from the issues discussed in 2.2 and 3.0 above, we initially based our analysis and design on the use of a fixed length packet in which the NLRAM information is accessible without the need to observe the contents of each packet at its underlying high bit rate (e.g., 40 Gbps). For the analysis below, we have used a packet length of 512 bytes, but longer or shorter packet lengths could be selected. The key issue to consider, in this section, is what method we will use to embed the NLRAM information in each packet. We considered several approaches, three of which are shown in figure 3.

In figure 3a we show an approach based on wavelength division multiplexing (WDM). The NLRAM information is carried on a second wavelength that is transmitted from router to router, over a common physical path (e.g., a fiber), with the rest of the corresponding packet. Through the use of WDM one can separate the NLRAM information from the rest of the packet using a wavelength demultiplexer. Similarly, one can add NLRAM information to a packet using a wavelength multiplexer. Since the number of NLRAM bytes is assumed to be much less than the total number of bytes within a packet, the data rate associated with the wavelength carrying the NLRAM information is much lower than the data rate associated with the rest of the packet.

In figure 3b we show an approach that is similar to that described above, but where the NLRAM information is superimposed upon each packet by modulating the power level of the packet envelope (i.e., its short term average power, averaged over a relatively large

number of bits). If the packet, excluding the NLRAM information , has a short term averaged balance of optical "1s" (pulse present) and optical "0's" (pulse mostly absent), then one could pass the packet through a modulator to impose the relatively slow NLRAM modulation of the average power level of the packet. If the packet, excluding the NLRAM information does not have a good short term averaged balance of 1's and 0's, then this approach is problematic.

In figure 3c we show an approach which utilizes a repetition of optical 1's to represent a logical "l" of the NLRAM information; and a repetition of optical 0's to represent a logical "0" of the NLRAM information. As a result, an optical receiver subsystem, whose purpose is to extract the NLRAM information, need only operate at the slower speed of B/N; where B is the underlying packet bit rate, and N is the number of times that optical 1's and 0's are repeated to represent the NLRAM 1's and 0's.

Other possible approaches, not shown in figure 3, include: the use of subcarrier modulation (of a subcarrier whose frequency is higher than the underlying packet bit rate) to superimpose the NLRAM information; or simply making the NLRAM bytes equivalent to any other bytes in the packet (which means that the packet needs to be examined at its underlying bit rate in order to extract the NLRAM information).

All of the approaches, described above appear to be viable because access to the NLRAM information is needed only at the input of the switching fabric (to read it) and the output of the switching fabric (if necessary to change it). Thus, the number of NLRAM (read/write) subassemblies is equal to, at most, the total number of router input and output ports. NLRAM subassemblies are not required elsewhere in the switching fabric. In addition, regenerative functions required at the input and output of the router include much of the circuitry (e.g. bit clock recovery and bit-value decision-circuitry) required to read NLRAM information from the high speed packet stream.

In the discussions below, we will show how several of these alternative approaches could be utilized.

A second consideration in the design of the packet structure (some might refer to this as the "layer 2 frame structure") is the incorporation of a mechanism that will allow packets to pass through the router switching fabric on a packet-by-packet basis without requiring the switching times within the router switching fabric to be comparable to the spacing between optical pulses in the packets (25 ps @ 40 Gbps). To accomplish this, we propose that packets will be separated by "guard bands" as shown in figure 4. Our requirement will be to keep the packets passing through the switching fabric separated in time (i.e., to retain the nominal guard band spacing to the extent possible).

4.2 Separation of Network Layer Router Address and Management (NLRAM) Information from the Incoming Packets

For each method of embedding NLRAM information into a packet, described in section 4.1, there is a corresponding method of separating or extracting the NLRAM information at the router input ports.

Figure 5a. shows how the NLRAM information can be separated from the remainder of a packet when the two-wavelength WDM method is employed. A WDM demultiplexer directs the wavelength carrying the NLRAM information to an NLRAM optical receiver ; while the remaining wavelength, carrying the rest of the packet, is directed to a full-data-rate optical regenerator. The WDM demultiplexer will introduce some attenuation in both intended paths; but this can be less than 1 dB. The WDM will also allow a portion of each wavelength to leak into the unintended path (crosstalk); but this is not critical, because each path through the demultiplexer is immediately followed by a receiver or a regenerator that will function properly (low error rate) if this leakage is less than 5%*.

*We select 5%, for illustration, to indicate that the leakage is not critical. Even larger leakage levels can be tolerated with proper design of the full-bit-rate regenerator and the NLRAM receiver, because of the large difference in the modulation rates of the two optical signals.

Figure 5b shows how the NLRAM information can be extracted from a packet when superimposed envelope modulation is employed to impress the NLRAM on to the packet. A beam splitter directs a portion of the incoming optical signal to an NLRAM optical receiver, while the remaining portion is directed to a full-data-rate optical regenerator. The beam splitter will introduce a loss in each path determined by the power splitting ratio, plus additional insertion losses. Since the modulation rate (bit rate) of the superimposed NLRAM information is considerably lower than that of the underlying packet bit stream, the (modulated) power level required for low error rate regeneration at the NLRAM receiver is proportionately lower than that required at the full-bit-rate regenerator. As a result, the fraction of the incoming optical power that must be directed to the NLRAM receiver will typically be less than 10%; depending upon the modulation depth of the superimposed NLRAM information, and the ratio of the NLRAM date rate to the underlying full data rate of the packet. This corresponds to a fractional dB reduction in the signal directed to the full-data-rate regenerator. The additional insertion loss of the beam splitter can be less than 1dB. If the NLRAM receiver is properly designed, it will be insensitive to the underlying high-bit-rate modulation; i.e., it will respond only to the modulation of the envelope of the optical power corresponding to the NLRAM information. Likewise, if the full-bit-rate receiver is properly designed, and the superimposed (NLRAM information) modulation depth is not too large (e.g., <10%), then the high-bit-rate regenerator will experience only about 1 dB of sensitivity penalty as a result of the superimposed modulation. The full-data-rate regenerator will remove all traces of the superimposed NLRAM modulation from the packet it regenerates.

Figure 5c shows how the NLRAM information can be extracted from a packet in which this information is part of the underlying packet bit stream. This approach is applicable whether the method of bit repetition described in 4.1 is utilized, or whether the NLRAM information is embedded in the packet at the normal packet data rate. The high speed

optical receiver acts both as a regenerator as well as an access point for observing the bit stream of each received packet. As discussed in 4.1, the high speed optical receiver includes a bit-rate clock recovery mechanism and a bit-value determination circuit (decision circuit) which provide a portion of the functionality needed to examine the passing bit stream to extract the NLRAM information at the full bit rate. However, some additional high speed circuitry is needed to perform this extraction. When the method of repeated bits is used to embed the NLRAM information in the high speed packet bit stream, then a receiver as shown in Figure 5d represents an alternative for avoiding this additional high speed extraction circuitry.

4.3 Regeneration to Remove the Effects of Transmission of Packets from their Sources

As described in 4.2, figures 5a. – 5d. all include a high bit rate (running at the packet's underlying bit rate) regenerator. The principal purpose of this regenerator is to remove the effects of transmission of the incoming packets from their respective sources. One can remove the effect of attenuation using an optical amplifier, and it may be possible to implement some of the other functions of a regenerator (clock recovery, sampling, comparing the sampled signal to a threshold each bit interval, and creating new optical pulses) using photonic devices of various kinds. While research results have been reported for "all optical" regenerators, a traditional regenerator, consisting of a high speed optical receiver and electronic clock recovery and decision circuitry, is assumed in this study. While traditional high speed regenerators running at data rates of 40 Gbps stretch the capabilities of electronic technologies, they are under development in R&D laboratories around the world for the next generation of SONET fiber optic transmission systems. Fortunately, as will be described below, these regenerators (and their asynchronous counterparts) will only be required at the input and output ports of the router.

4.4 Aligning Arriving Packets with the Router Master Clock

Packets arriving at the router from different sources will experience different amounts of delay as they travel the associated distances through fiber or "free space" media. In addition, we should assume that each source of packets has its own local master clock; and that each of these is not perfectly synchronized with any other local master clock in the network. We can, however, assume that each local master clock is accurate to within a small percentage (e.g., <1 ppm) with respect to a nominal clock frequency. To assume that all local master clocks are synchronized would place a severe constraint on the overall network architecture, and would substantially increase the difficulty of implementing that architecture.

Figure 6 shows a design for implementing a "packet aligner" that would allow incoming fixed-length packet start times to be aligned with the local master packet clock of the (mostly) photonic router switching fabric. This design builds on the assumption that we are receiving fixed length packets; and that there is a guard band between packets. For illustration in this report, we assume fixed length packets containing 512 bytes (4096

bits). The bit rate is nominally 40 Gbps. The length of a packet is nominally 25 ps/bit x 4096 bits/packet ~ 100 ns. We assume that the guard bands are 10% of the packet length, or ~ 10 ns. Thus the spacing between packets is ~ 110 ns. We propose to design the switching fabric to utilize switching transition times of ~1 ns (10% of the guard band duration); and we intend to align the starting times of packets relative to the local master packet clock to within 1ns, at each packet aligner used within the fabric.

In figure 6, we have a sequence of seven (7) 2 x 2 photonic switches, each of which can introduce a fiber delay line into the path of an optical packet through the aligner. In each stage, the delay is either d (a nominal delay associated with the interconnections between adjacent switches and the delays through the switches themselves) or d + [(1 ns) x 2**(n-1)]; where n is the stage number. Thus the selectable delays, not including the fixed delay d, are: 1 ns, 2 ns, 4 ns, 8 ns, 16 ns, 32 ns, and 64 ns. By inserting (or not inserting) these selectable delays, we can produce a total delay through the packet aligner of between 7d and 7d + 127 ns, in steps of 1 ns.

Also shown in figure 6 is circuitry for adding optical amplification to compensate for the insertion loss of the packet aligner and to control the value of delay inserted by the aligner.

Before proceeding to describe this circuitry and the necessity for including its various components, it is helpful to estimate the following parameters for the 7 stage aligner delay subassembly: nominal loss, statistical loss variability from nominal loss, nominal delay, statistical delay variability, and crosstalk.

For the purposes of this report, we shall assume the following characteristics for the 2 x 2 switching elements (crosspoints), their interconnections, and the delay lines:

Nominal switch insertion loss of one 2 x 2 switching element: 1.0 dB
Statistical insertion loss variability of one 2 x 2 switching element: +/- 0.5 dB
Nominal coupling loss in or out of one 2 x 2 switching element: 0.2 dB
Statistical coupling loss variability: +/- 0.1 dB
Nominal delay through one 2 x 2 switching element: a known design parameter (value not important)
Statistical delay variability through one 2 x 2 switching element: < +/- 25 ps
Statistical deviation of a fiber delay line's delay from nominal: < +/- 25 ps (+/- 0.5 cm)
Worst case crosstalk in one 2x2 switching element: -30 dB (optical power ratio)

Using the above parameter values, we obtain the following for the 7 stage packet aligner:

Nominal insertion loss: 9.8 dB
Statistical insertion loss range: 4.9 – 14.7 dB
Statistical delay deviation range: +/- 350 ps
Worst case crosstalk (assuming power summation): <-43.7 dB (optical power ratio)*

*In order for an optical packet to interfere with itself, or with subsequent packets passing through the packet aligner, there must be two (2) passes of the power through the 30 dB of attenuation associated with a crosstalk path. This results in 60 dB of attenuation for every such exposure. There are 42 independent double-crosstalk exposure paths in a 7 stage aligner. The optical packet can be coupled by crosstalk into the wrong path by any of the seven 2 x 2 switches. After that, this coupled power (more than 30 dB below the optical packet power level) can crosstalk back into the optical pulse (another 30 dB of attenuation) in any subsequent stage. The total number of such double crosstalk paths is $6+5+4+3+2+1 = 42$. $60\ \text{dB} - 10 \log(42) = 43.77\ \text{dB}$

Returning now to the aligner control components:

With a range of 4.9-14.7 dB, the worst case insertion loss of the aligner can be more than 50% of the allowable loss; before gain (or regeneration) is required to prevent the optical packet from falling below the level from which the underlying information can be recovered with a low error rate. With a loss variability of +/- 4.9 dB, there is some concern regarding the susceptibility to crosstalk in the main switching matrix (to be described below) that optical packet streams arriving from different input ports might experience. I.e., a lower power optical stream is susceptible to crosstalk from a higher power stream. Since the next stages of the switching fabric (leading to the packet buffers) will introduce additional nominal loss and additional statistical loss variability, one might decide to wait before adding gain to compensate for the loss of the aligner. However, for reasons that will be evident in the later sections, we shall assume that an optical amplifier is used, as shown in figure 6, to compensate for the aligner loss. We assume that a beam splitter is inserted after the optical amplifier, to split off a portion of the optical signal. This split portion of the optical signal enters a receiver module whose purpose is to measure the peak power level of the amplifier output signal, and to determine the leading edge of the packet power envelope with an accuracy of +/- 1 ns. These receiver outputs are used in separate feedback loops to adjust the amplifier output power level to nominal, and to select the proper value for the aligner delay.

4.5 Packets Which Are Competing for the Same Output Port

One of the principal functions of the router switching fabric is to buffer (temporarily store) packets which are destined for the same output port. This buffering can occur either at the input of the switching fabric or at the output of the switching fabric. In this report we shall assume that the buffering occurs at the input of the switching fabric.

We will assume that we require enough buffering capability to temporarily store each incoming packet for between 0-31 packet durations (actually the spacing between packets, which includes the guard band).

An important issue will be whether or not we can compute, upon the arrival of a packet, the specific number of packet intervals for which it will need buffer storage. If we assume that we are employing a "first-come first-served" buffering algorithm, where each packet destined for a given output port, and requiring the use of a given input port, must only

wait for packets which have arrived earlier and which are already in queue for those same ports, then we can employ buffer design "A" described below. However, if we assume that packets in queue for a given output port can be preempted by packets which arrive later, but which have a higher priority for access to the input or output port, then we cannot use buffer design "A", and must revert to buffer design "B".

Buffer design A is shown in figure 7a. It is very similar to the packet aligner described in 4.4, above. Each of the five (5) stages of the buffer consists of a 2 x 2 switching element, plus a delay line. The delay line values are (not including the "0" delay nominal value of "d" nanoseconds) : 1, 2, 4, 8, and 16 packet intervals. If a packet interval (including the guard band) is ~ 110 ns, then the shortest fiber delay line is approximately 22 meters long. The longest delay line is approximately 352 meters long. Using this multistage delay line, we can insert predetermined delays of between 5d and 31T + 5d where T is the packet spacing. Using the same parameter values as in the packet aligner, and adding a new parameter for fiber loss, we obtain the following:

2 x 2 switching element insertion loss (including input and output losses): 0.7 – 2.1 dB
Delay uncertainty for one stage +/- 50 ps
Crosstalk in one 2 x 2 switching element <-30 dB (optical power)
Fiber nominal loss < 0.5 dB/km

For a five (5) stage buffer, we obtain the following parameter ranges:

Insertion loss: 3.5 – 10.85 dB (including a maximum of 682 meters of fiber)
Crosstalk* : -50 dB [ten (10) 2-pass crosstalk paths @ -60 dB each]
Statistical delay deviation from nominal (jitter): +/- 250 ps

*Crosstalk of a packet into itself or subsequent packets

In order to provide access to buffer storage for all packets, in the event that each packet in a 31 packet sequence requires 31 packet-durations of buffer storage, we propose to provide 32 buffers for each router input port. A more detailed, future analysis of the buffer requirements will probably suggest that this worst case design is not optimal. I.e., it may be possible to obtain essentially the same overall router performance with fewer buffers per port; taking into account the likelihood that all of the packets in a contiguous 31 packet sequence will not require 31 packet-durations of buffer storage, and the possibility that two packets can share the same buffer at the same time for some combinations of required buffering.

To provide 32 buffers, we require a 1 x 32 switching fabric between the output of each packet aligner and the set of 32 buffers. We will discuss this further below.

Buffer design B is shown in figure 7b. In this design, buffer delays are added in fixed increments of the spacing between packets. To obtain delays as large as 31 packet intervals, without using 32 stages of 2 x 2 switches and delay lines, we utilize a feedback

design. To maintain a constant nominal delay through the buffer, regardless of the number of passes through the feedpack path, we introduce a shortened delay in the feedback path; in this example T-(Nd + a). The factor (Nd + a) corresponds to the 1-pass delay through the N buffer stages (Nd) + the delay optical amplifier and the last 2x2 switching element (a). Note that we place the amplifier in the single-pass chain to avoid having to adjust its gain each time the number of required loops around the single pass chain changes.

Using the same parameter values as before, we find that the range of losses (not including the gain or insertion loss of the optical amplifier and its associated 2 x 2 switching element) are as follows:

1-pass loss: (0.7N – 2.1N) + fiber loss (<0.34 dB)
Maximum loss (multiple pass to produce 31T delay) > 32 x 2.1 ~ 67.2 dB
Statistical delay deviation from nominal delay > +/- 1.6 ns

From the above, we observe that for buffer design B, the (non-amplified) buffer loss range is too large to be accommodated by the allowable loss budget (<30 dB). This demonstrates the need for an optical amplifier. In addition, the statistical delay deviation from nominal (jitter) is larger than the desired deviation (1 ns).

Thus, for design B, we require an optical amplifier for each buffer.

Rather than introduce a delay aligner for each buffer, we will allow the jitter to exceed the target range, and remove it later as described below. I.e., our target range of +/- 1 ns is fairly tight compared to the 10 ns guard band between packets, and we can afford to relax it somewhat.

A 32 stage buffer will introduce 496 [(32 x 31)/2] opportunities for the optical pulse to crosstalk into itself or subsequent packets; each of which involves the double crosstalk attenuation of >60 dB. Some of these exposures will be exacerbated by the possibility that the attenuation through a stage could be lower for the undesired (crosstalk) signal than the desired signal, thus lowering the contrast as both signals enter the next switching stage. On the other hand, the feedback design eliminates many of these exposures by breaking the path that carries the crosstalk signal.

Based on the above, the best value for N (to minimize crosstalk) is zero, but larger values of N (e.g., N= 8) will produce acceptably low crosstalk; and may make it easier to adjust and stabilize the gain around the optical feedback loop. I.e., a moderate value of loop loss (~ 16 dB) may be easier to compensate than a small loop loss (~ 2 dB).

With design B, it is possible to share a buffer among multiple packets. As packets pass through the design B buffer they should either be routed to an additional T second delay, or they should be routed to the buffer output (lower rail of the buffer). Thus each packet should either be delayed by T seconds or leave the buffer. Two packets injected (sequentially) into the buffer will not "catch up with" each other, and will not interfere

with each other; provided: we do not try to route two packets out of the buffer at the same time (which should never be done); and we do not inject a packet into the buffer at the same time that an existing buffered packet is returning to the input via the feedback loop.

As a result, an N stage buffer, of design B, can accommodate up to N independent packets simultaneously. [Note that a buffer of design A cannot accommodate multiple packets, because: packets can "catch up" to each other as various per-stage delays are selected; and packets may use the lower rail even when they are not exiting the buffer.

As mentioned above, we need a 1 x 32 switch to enable packets on a given input port to access each of 32 possible delay-line buffers. We will also need a 32 x 1 switch to redirect the outputs of those buffers to their associated single input port on the main switching matrix.

Figure 7c shows a 1 x 8 banyon switch, which also can serve as a 8 x 1 banyon switch when the input and outputs are reversed. The switch has 3 stages, each of which consists of a plane of 2 x 2 switching elements. By extending this concept to 5 planes, one obtains a 1 x 32 or a 32 x 1 banyon switch. Since only one packet will be passing through the switch per packet interval, there is no problem associated with blocking (two packets trying to use the same part of the switch at the same time).

Using the parameter values above for 2 x 2 switching elements, and their input and output losses, we obtain the following for either the 32 x 1 or the 1 x 32 switch:

1 x 32 switch insertion loss range: 3.5 – 10.5 dB
Delay deviation from nominal: +/- 250 ps
Crosstalk: no exposures

If we now combine the 1 x 32 switch, the buffers, and the 32 x 1 switch, we obtain the following

Case 1 Buffer design A

Total insertion loss: 10.5 – 31.85 dB
Delay deviation from nominal: +/- 750 ps
Crosstalk < -50 dB

Case 2 Buffer design B

Total Insertion loss: 7.0 – 21
Delay deviation from nominal: +/- 2100 ps
Crosstalk < -50 dB

For a buffer subsystem using buffer design A, the worst case loss (31.85 dB) will likely exceed the allowable loss budget. However, rather than assuming that we will insert an optical amplifier after each buffer (32 amplifiers), we will assume that we can control the

manufacturing process so that subsystems employing buffer design A have insertion losses that stay within the loss budget. Therefore we assume that only one optical amplifier, inserted after the 32 x 1 switch, is required.

For a buffer subsystem using buffer design A, we will insert both an optical amplifier and a 0-7 ns (3 stage) delay line packet aligner after the 32 x 1 switch to reduce the +/- 2100 ps of delay deviation. Since +/- 1600 ps of this delay deviation results from path-dependent delay through each of the buffers, this delay equalizer must measure the rising edge of the envelope of each packet that passes through it, and adjust the delay accordingly. Figure 7d shows the required optical amplifier and packet aligner subassembly.

4.6 The Main Switching Matrix

We shall utilize a rearrangably non-blocking Benes fabric for the main switching matrix. The Benes switching matrix concept is shown in figure 8, for an 8 x 8 switch with 5 planes of 2 x 2 switching elements (crosspoints) . The matrix is called a rearrangably non-blocking matrix because each path through the matrix from an input to an output must be computed simultaneously with all of the other paths (or pre-stored in a look-up table). That is, packets destined for different outputs of the combined matrix must not attempt to use the same matrix resources, at the same time, in a way that results in a collision. There are switching fabrics that are "strictly non-blocking". With such fabrics, a path through the fabric can be established from any input to any (idle) output, independent of existing paths through the fabric. However, the number of planes of 2x2 switching elements required to construct a strictly non-blocking N xN matrix is believed to grow as log N + (log N)**2. Thus a 128 x 128 strictly non-blocking matrix (e.g., a Batcher-banyon matrix) would require 7 + 49 = 56 planes of 2x2 switches. An 8x8 Batcher-banyon switching matrix which can independently switch 16 wavelengths in parallel (see discussion of wavelength-space wavelength switching below) would require 3 + 6 = 9 planes.

The Benes fabric we require must have 128 input ports and 128 output ports. It consists of 13 planes of 2 x 2 crosspoints. Using the same 2 x 2 crosspoint parameter values that we have used before, we obtain the following

Main switching fabric insertion loss: 9.1 –27.3 dB
Delay deviation from nominal (jitter) : +/- 650ps
Crosstalk (13 exposures @ -30 dB each : -18.8dB (assuming power addition, and not including increased crosstalk exposure between unequal-power-level packets)

As in the case of the buffer subsystem, we will require an amplifier or a regenerator at each output of the main switching matrix. We may require a plane of optical amplifiers between matrix component subsystems (128 amplifiers). Since we will require regenerators at the output ports of the router, it would be desirable to avoid amplifiers

directly at the output of the main switching fabric. The ability to do this will depend on the value of the allowable loss budget.

4.7 Asynchronous Regeneration of Packets

Each of the packets that arrive at an output port of the main switching matrix may come from any of the possible input ports of the router, will have experienced a statistically variable attenuation (through the main switching matrix), and will have a statistically variable position in its packet time slot relative to the nominal position. As a result, we will require asynchronous regenerators at the router output ports. A block diagram of an asynchronous regenerator is shown in figure 9.

Each asynchronous regenerator contains of an optical detector/preamplfier/variable-gain-amplifier chain to convert the incoming packet bit streams into electrical bit streams of suitable amplitude. Because the power level of each packet will vary from packet- to-packet, a fast acting automatic gain control (AGC) system is required.

After the variable gain amplifier chain, the electrical pulses pass through a comparitor which determines whether each pulse is above or below a selected decision threshold. The output of the comparitor must be sampled to make bit-by-bit decisions as to whether each optical pulse is on or off. One needs a sampling clock which is optimally (or close to optimally positioned in time relative to the electrical pulse stream in order to make these decisions with a low probability of error.

A local bit-rate-clock drives a phase shifter that generates N outputs, each of which is 360/N degrees out of phase with its neighbors.

A phase selector attempts to select the best bit-rate-clock output phase to use for sampling the output of the comparitor. It does this by performing nonlinear operations on combinations of the electrical bit stream (amplifier output) and the phase-shifted clock outputs.

Having selected a clock phase, this phase-shifted clock output is used to sample the comparitor output and regenerate (recreate) optical pulses. It is assumed that each packet bit stream has an associated clock that is within +/- 1 ppm of a nominal clock rate, and that the local bit rate clock also is within this tolerance. Thus, if the selected clock phase is acceptable for sampling the comparitor output, it will stay within the acceptable range for the 4096 bits intervals of the packet.

Asynchronous regenerators have been used for decades at lower data rates (e.g., in the UARTS associated with many computer links), and are being designed for 40 Gbps operation (ref: OFC '99).

An important design question is whether or not the asynchronous regenerator output packets should be electronically buffered (in a high speed RAM) to produce a

synchronous output signal from the router. Such a synchronous output signal would be easier to accept as an input to a downstream router or other termination.

4.8 Adding NLRAM Information to the Outgoing Packets

If the NLRAM information is incorporated into the packets using 2-wavelength WDM, then the subassembly shown in figure 10a can be used to add this information to the outgoing packets.

If the NLRAM information is incorporated into the packets using envelope modulation, then the subassembly shown in figure 10b can be used to add this information to the outgoing packets.

If the NLRAM information is incorporated directly into the bit stream of the packet, then the subassembly shown in figure 10c can be used to add this information to the outgoing packet. The subassembly used in figure 10c can also be combined with the circuitry (random access memory) required to buffer packets to create a synchronized output stream as described in 4.7.

4.9 Putting It All Together

Figure 11 shows the complete router switching fabric design (using the 2-wavelength method for incorporating NLRAM information).

Packets arriving at a router input port first pass through a WDM demultiplexer; in which the NLRAM information is stripped off. The remainder of each packet then enters a regenerator; which removes the effects of transmission from the distant source. Regenerated packets pass through a packet aligner; which aligns the packet stream with the local packet clock. The aligner also includes an optical amplifier to compensate for the insertion loss of the aligner.

Packets then pass through a buffer subassembly consisting of: a 1 x 32 banyon switch; 32 available, 0-31 packet, delay buffers; and a 32 x 1 banyon switch. Depending upon the design of the 0-31 packet delay buffers, each of these buffers may incorporate an optical amplifier. The output of the buffer subassembly includes an optical amplifier.

Packets then enter the 128 x 128 main switching matrix (a Benes matrix with 13 stages). The main switching matrix may require a plane of optical amplifiers between two sub matrices to compensate for accumulating insertion losses.

The outputs of the main switching matrix direct packets to a set of 128 asynchronous regenerators. After each regenerator, NLRAM information is added using a wavelength multiplexer.

4.10 Performance of the Switching Fabric

We have analyzed (with some approximations) the characteristics of offered load vs. the probability of buffer overflow (packet loss) using the standard statistical methods that have been applied to the analysis of packet switches. Figure 12 shows, for various values of utilization, i.e., the throughput, as a percentage of the (unachievable) nominal aggregate capacity of the router (maximum packets per second per port x # of ports), the probability of buffer overflow for various buffer sizes. We see that our example of a 128 port router, capable of buffering up to 32 packets per port, will achieve a very low probability of buffer overload at up to a 50% value of utilization (U).

It should be noted that traditional statistical models, which are used to analyze buffer size requirements (as a function of utilization and packet loss), tend to underestimate the required buffer sizes in real applications. This is because real packet traffic has been shown to have "self-similar" behavior; which implies that long sequences of packets, destined for the same output port, are more likely to occur than traditional models predict. Our analysis is based on traditional statistical models. Further analysis would be required to employ self-similar traffic models.

4.11 Observations

The conceptual design process, described above, did not uncover any "showstoppers" that would prevent the implementation of a first generation router switching fabric based on (mostly) photonic devices. The objective for 1 ns of statistical deviation of a packet's position relative to the local packet clock probably should be loosened, so that packet aligners are not required, except at the input ports of the router. With a nominal 10 ns guard band, a deviation of +/- 5 ns should be quite acceptable. At the assumed level of –30dB of crosstalk in each 2 x 2 switching element, and assuming that crosstalk accumulates on a power basis, crosstalk was not a problem. The worst case cumulative loss through the subassemblies necessitated the use of optical amplifiers. With closer tolerances on subassemblies, to reduce the statistical loss variations through these subassemblies, it may be possible to use fixed gain amplifiers to compensate for insertion losses (rather than amplifiers which must adjust their gain on a packet by packet basis).

5.0 Second Generation Design: A Wavelength-Space-Wavelength Switch

We explored the potential benefits and the associated challenges of a second generation design based on advanced wavelength multiplexing technologies.

Specifically, we considered a wavelength-space-wavelength (W-S-W) switch as shown in figure 13.

The name "wavelength-space-wavelength switch" is selected by analogy to the time-space-time switches used in digital switching systems. In these conventional circuit switches, a space switch is "reused" in each of N time slots that make up a repeating "frame" sequence.

At the input of the W-S-W we show 8 wavelength multiplexed port-groups; each of which contains 16 packet streams on 16 corresponding wavelengths. Thus the total number of equivalent input ports is 128.

In the wavelength-space-wavelength switch, eight (8) "wavelength interchanger" subsystems allow each incoming packet to be moved to a different wavelength; provided that within each port-group of 16 wavelengths, no two packets can occupy the same wavelength at the same time. As is well known in the optical networking field, implementing wavelength translators is a very challenging device challenge.

The packets that leave their respective wavelength interchangers enter a pair of 8 x 8 "wavelength-divided" space switches. These are analogous to the time-divided space switches in a time-space-time switch. The use of two 8 x 8 switches ensures that the wavelength-divided space switch will be non-blocking. This result derives, by analogy, from the existing theory of time-space-time switches.

The wavelength divided space switch must provide an independently controllable connectivity pattern for each separate wavelength. Thus the 2 x 2 crosspoints that make up this matrix must provide independently selectable switching for each wavelength (simultaneously). This is another very challenging device challenge.

The advantage of a W-S-W architecture is in the reduction of the number of components required to implement the main switching matrix. This implies a potential reduction in the cumulative loss, jitter, and crosstalk of the matrix (fewer planes of 2 x 2 switching elements); and may imply a reduction in the size and cost of the matrix (fewer parts, fewer interconnections).

6.0 Networking Issues

As discussed in 2.0, there are a number of networking-related issues that must be addressed in the design of a router switching fabric based on (mostly) photonic components.

In our design, in 4.0 above, we have not included any discussion of the SONET layer. The packets were assumed to arrive periodically, directly on top of the optical layer.

It would be possible to utilize a SONET frame structure in our proposed design. In this case, the IP packets (with or without NLRAM information) could be inserted within the SONET payload. The regenerator at the input of our router (Figure 10) would have to be augmented with additional circuitry. This would include a random access memory for packet buffering, and circuitry to separate the packets from the SONET frame. The embedded network management information would be recovered from the overhead fields of the SONET frame, and the packets would be delivered to the switching fabric with a uniform packet spacing. Likewise, the output ports of the router would have to include circuitry for creating the SONET frame sequence, incorporating network

management information into the overhead fields, and for inserting packets into the SONET frames.

As an alternative, we could employ an IP-over-optical layer approach (no SONET layer). In this case, as discussed in 2.1 we would need a mechanism for conveying network management information (e.g., for establishing "paths"). This can be accomplished, as it is in IP networks today, by utilizing network management packets to convey network management information between routers. In addition, we can utilize the NLRAM information to convey all required network management information.

To recreate paths between selected endpoints, we can reserve capacity through routers for different "virtual paths". These virtual paths would be represented by a portion of the address field associated with each packet, or equivalently, another field within the NLRAM information. If the routers manage traffic in compliance with these reservations, we can create the equivalent of point to point paths through a multi-router network.

There will also be networking issues associated with the use of large, fixed length packets. While the use of such packets does not appear to present serious problems in high capacity backbone router applications, the full implications of the use of such packets requires further study.

7.0 Conclusions

We conclude that, using existing technologies, it is feasible to implement an ultra high capacity (1 billion packets per second) router switching fabric: utilizing (mostly) photonic components. Such a router holds out the promise of significantly reduced physical design challenges, compared to traditional electronic router switching fabrics. We also conclude that advances in optical networking technologies (wavelength translators and wavelength selective crosspoints) would enable the implementation of a wavelength-space-wavelength W-S-W) switching fabric. A W-S-W switching fabric would require a substantially reduced number of optical components and interconnections. We see no major obstacles, related to implementing network management functionality, which would be introduced by the proposed use of long, fixed length packets, and the absence of a SONET layer.

Figure 1

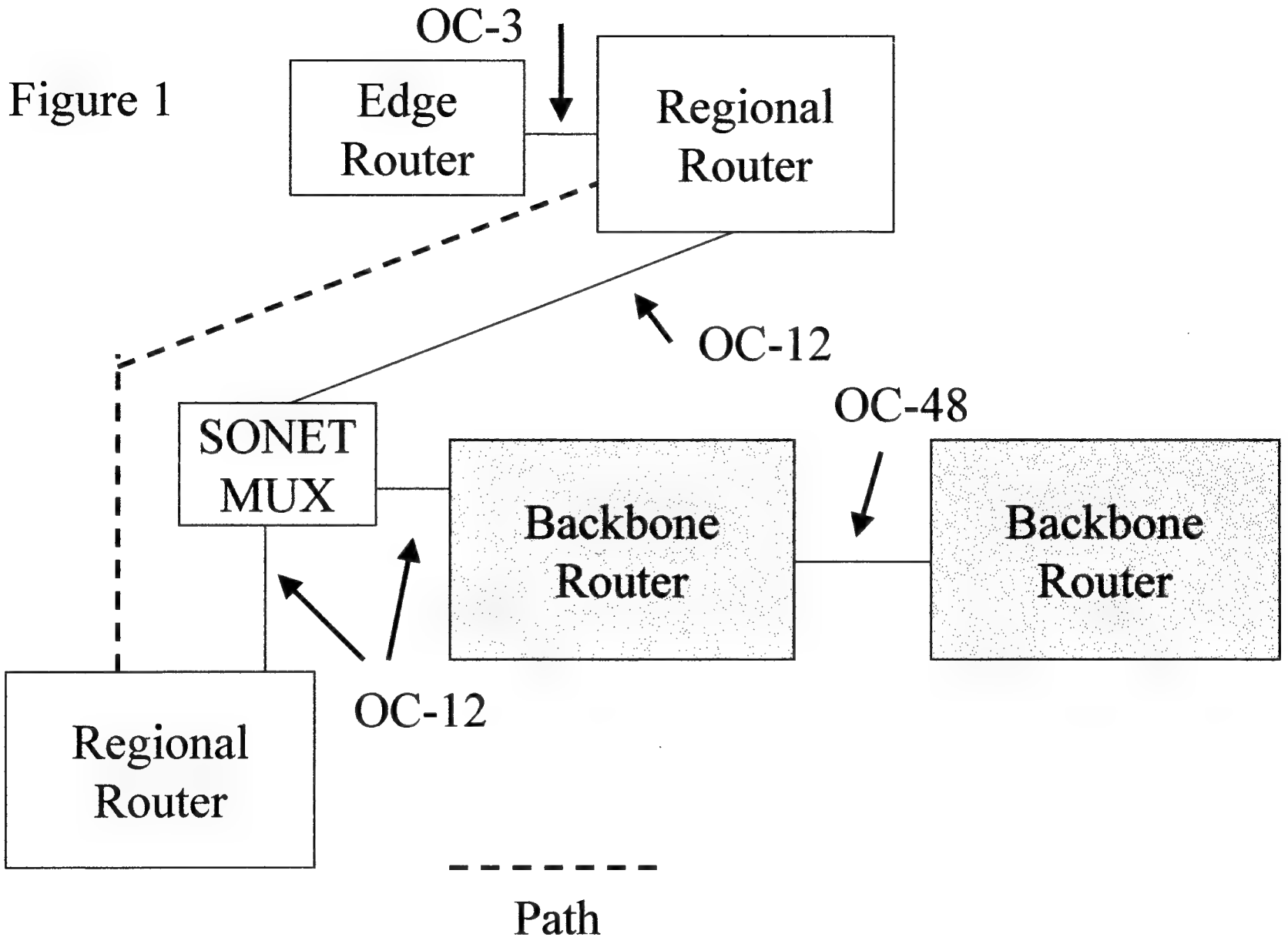


Figure 2 Separate mechanisms for placing NLRAM and payload information on the optical layer (violates traditional protocol stack paradigm)

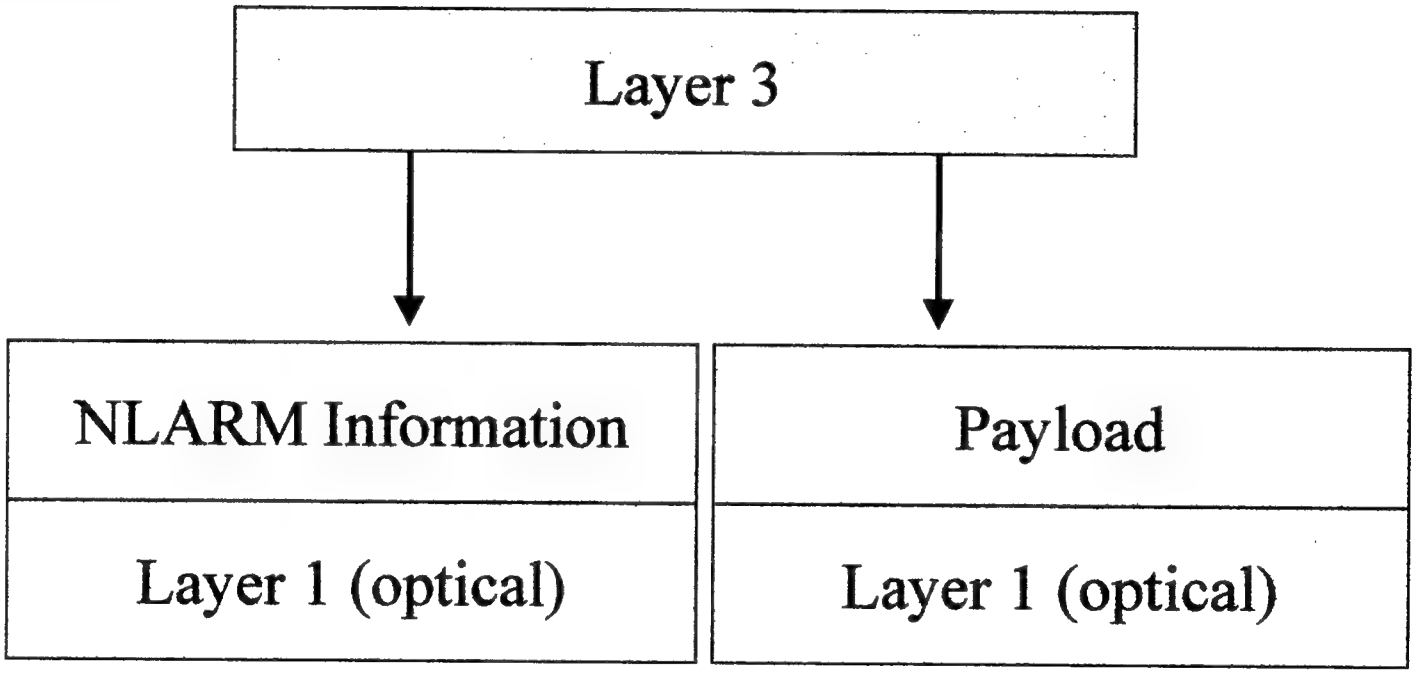


Figure 3a Inserting NLRAM Information: WDM

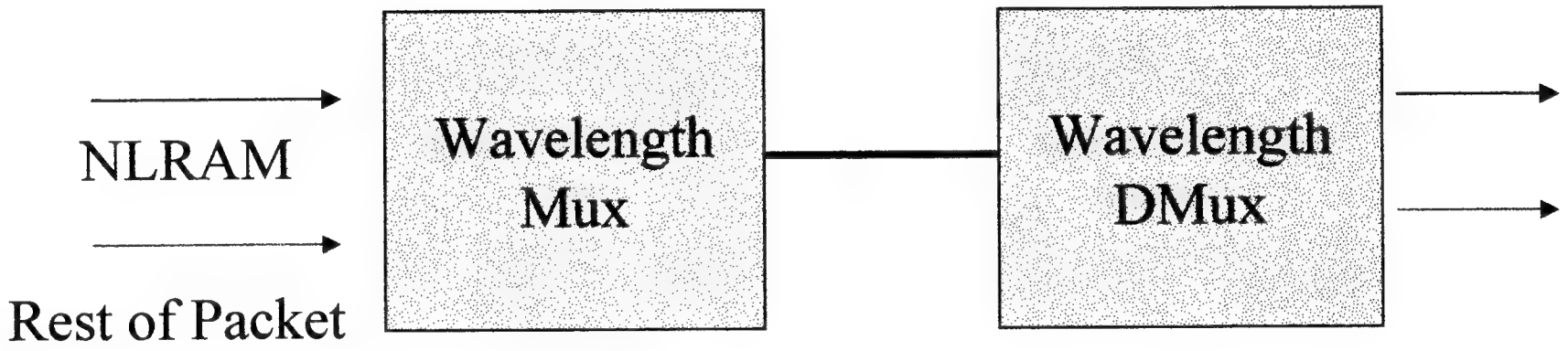


Figure 3b Inserting NLRAM Information:
Overlay modulation

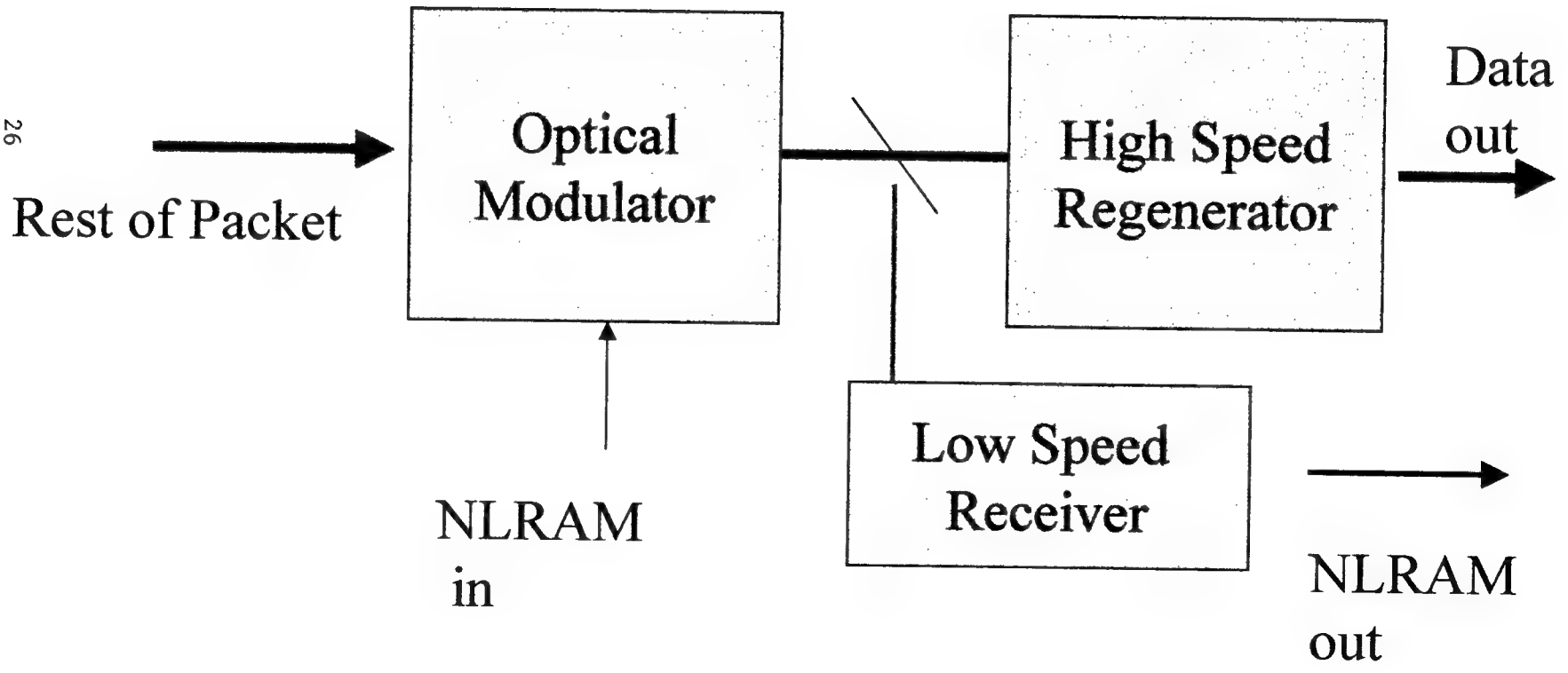
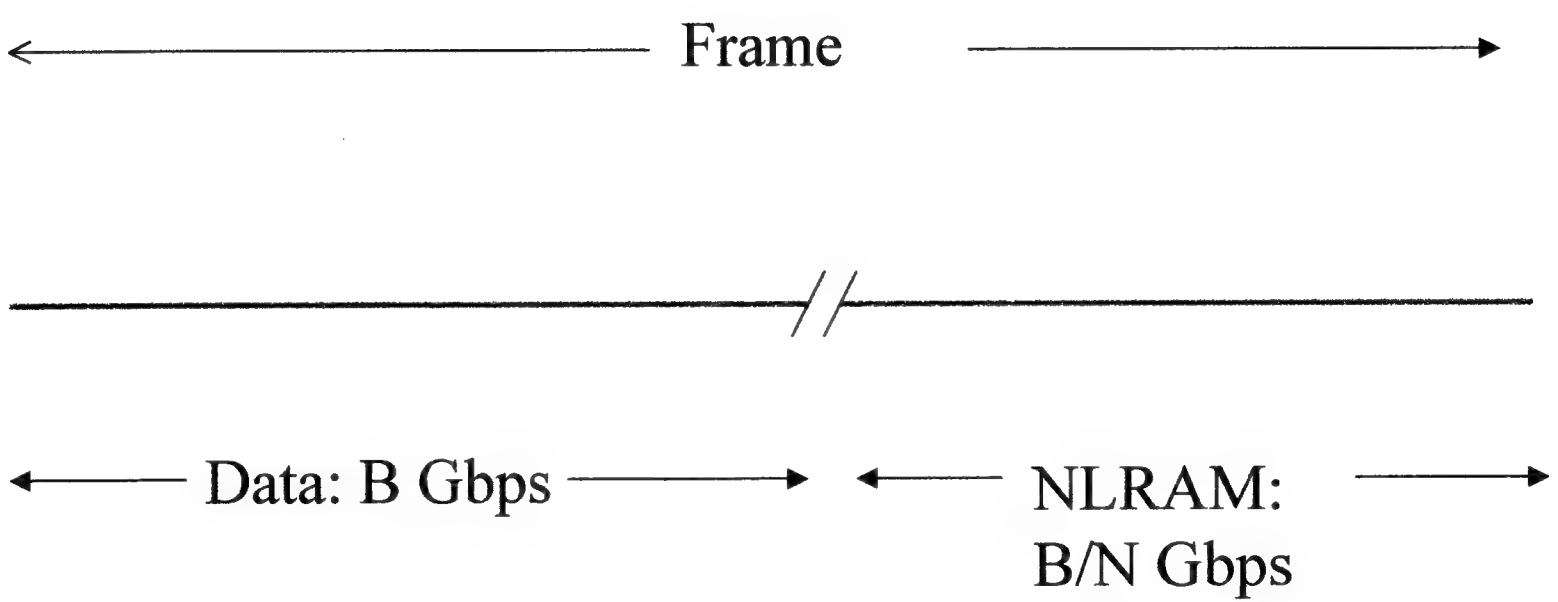


Figure 3c Inserting NLRAM Information: Repeated bits



N= the number of high speed bits that are repeated to construct one NLRAM bit

Figure 4 Guard band between optical packets

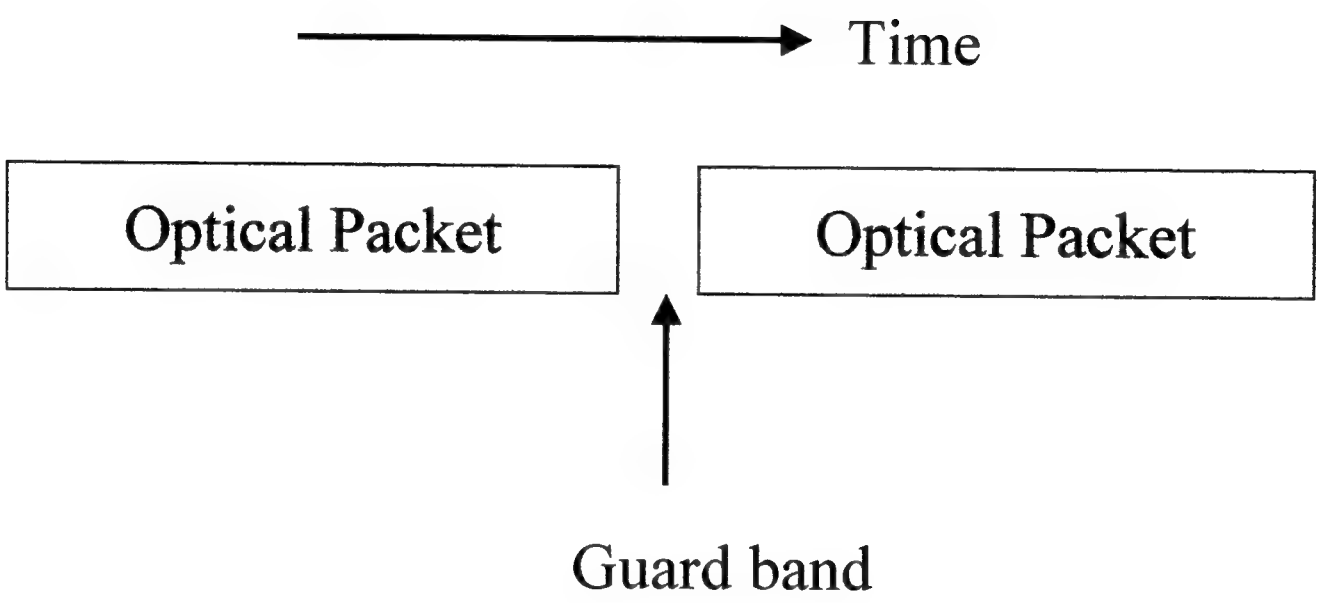


Figure 5a Separating NLRAM: WDM

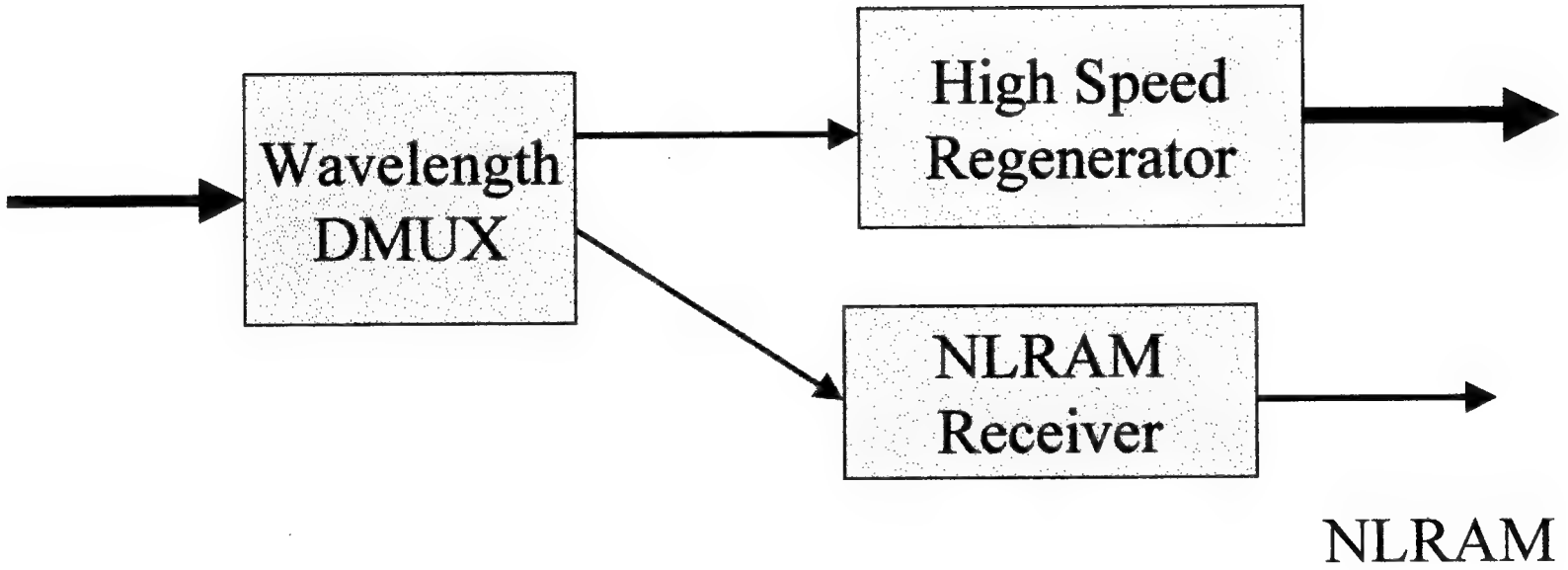


Figure 5b Separating NLRAM: Overlay modulation

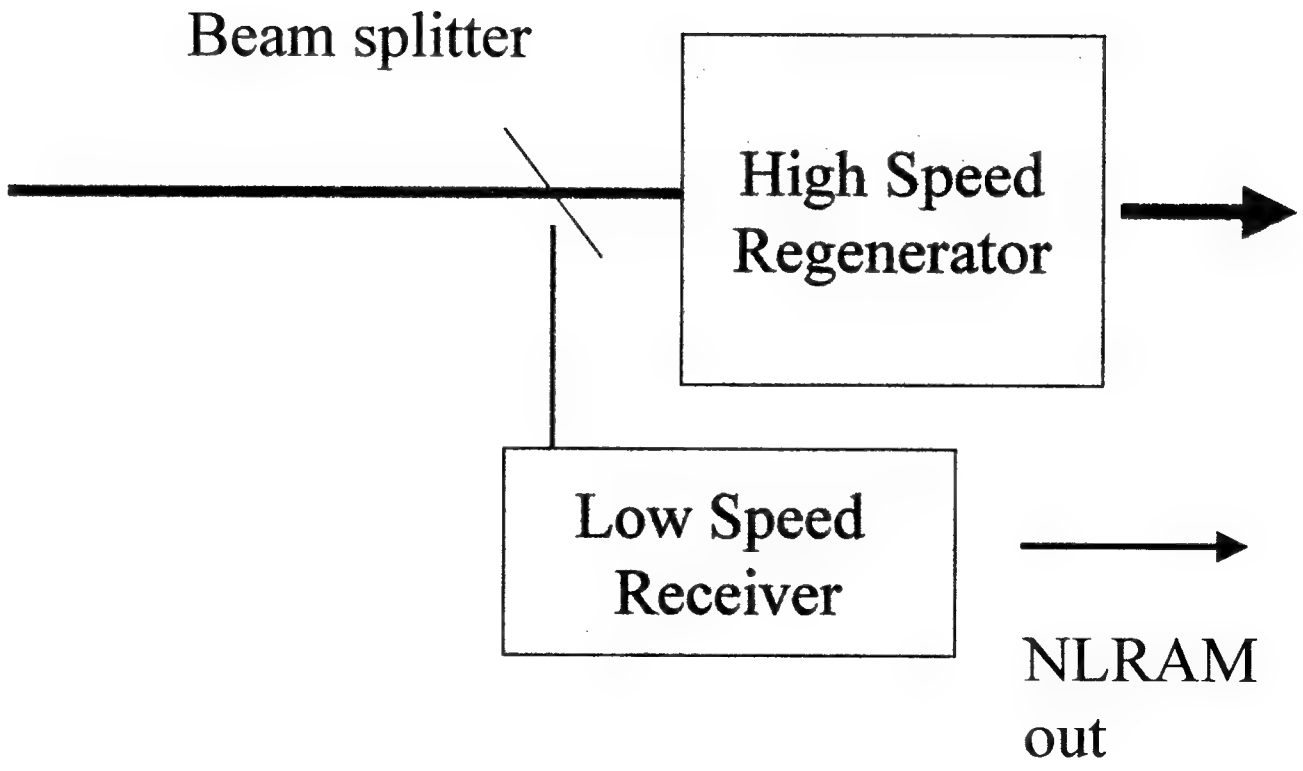


Figure 5c Separating NRLAM: Embedded in packet stream

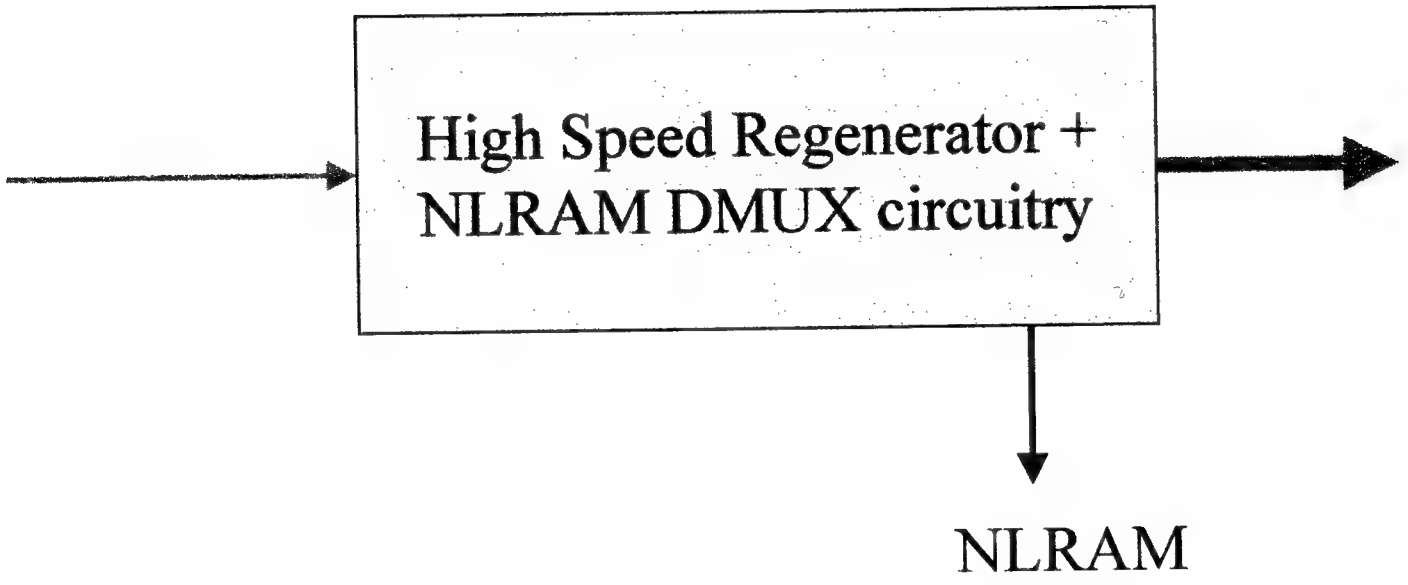
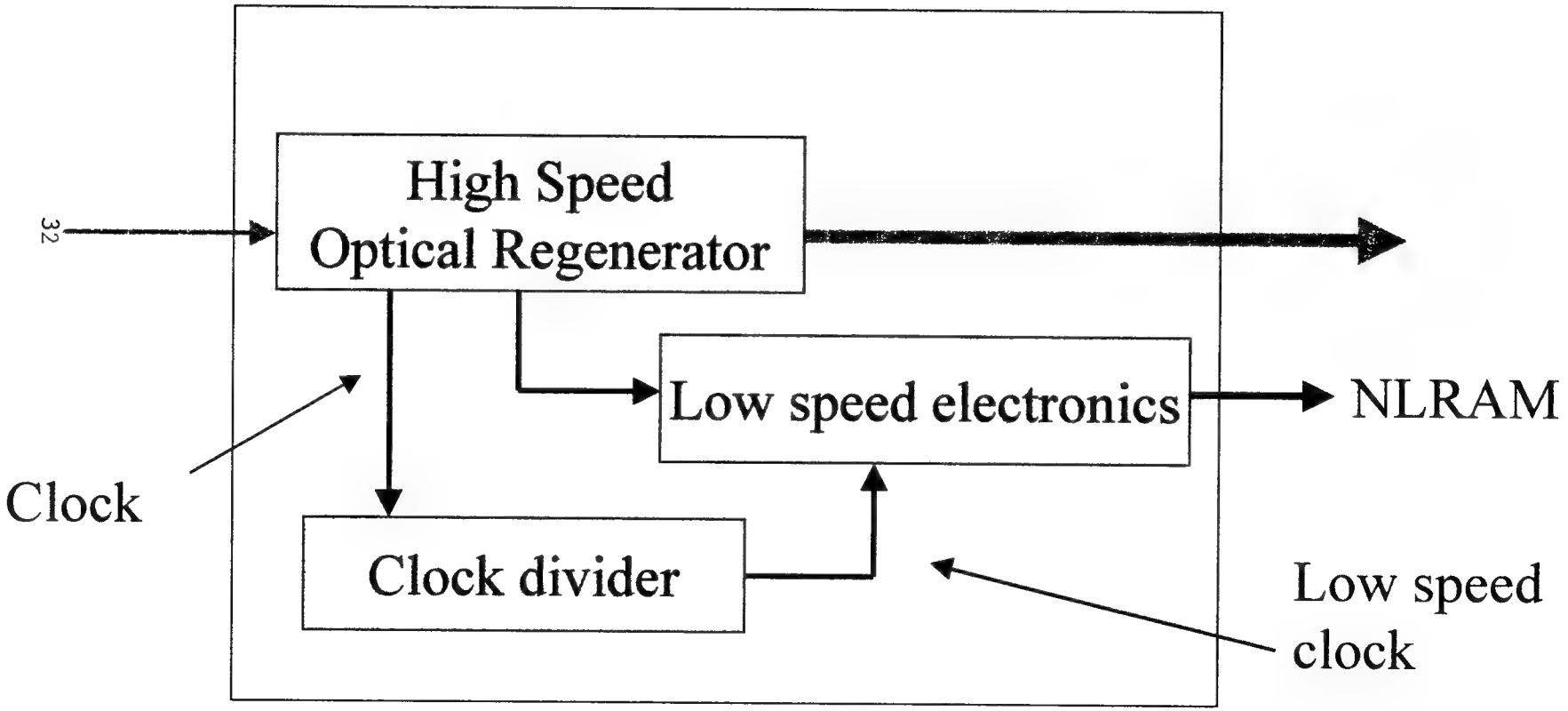


Figure 5d Separating NLRAM: Embedded in packet stream



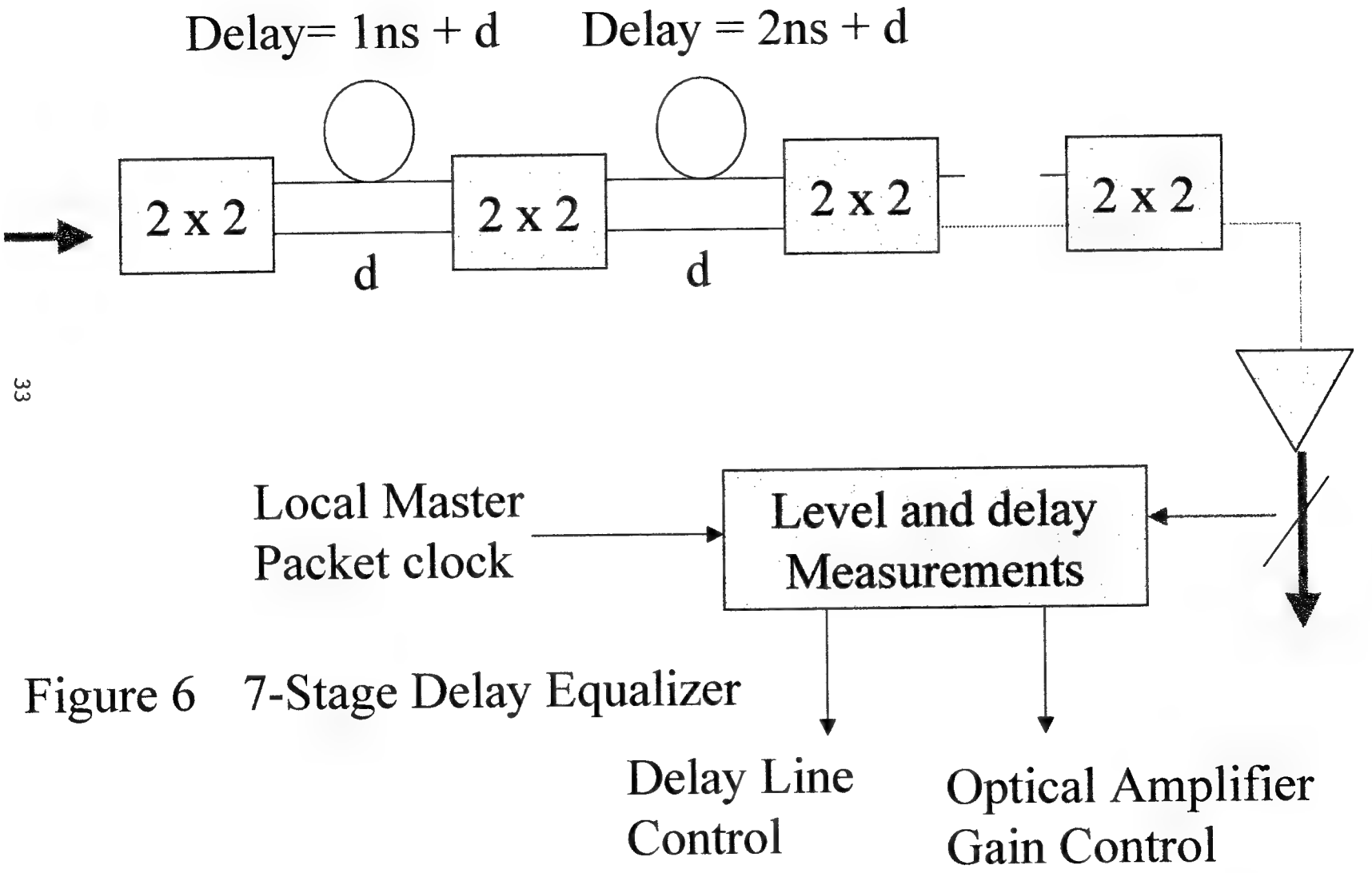


Figure 6 7-Stage Delay Equalizer

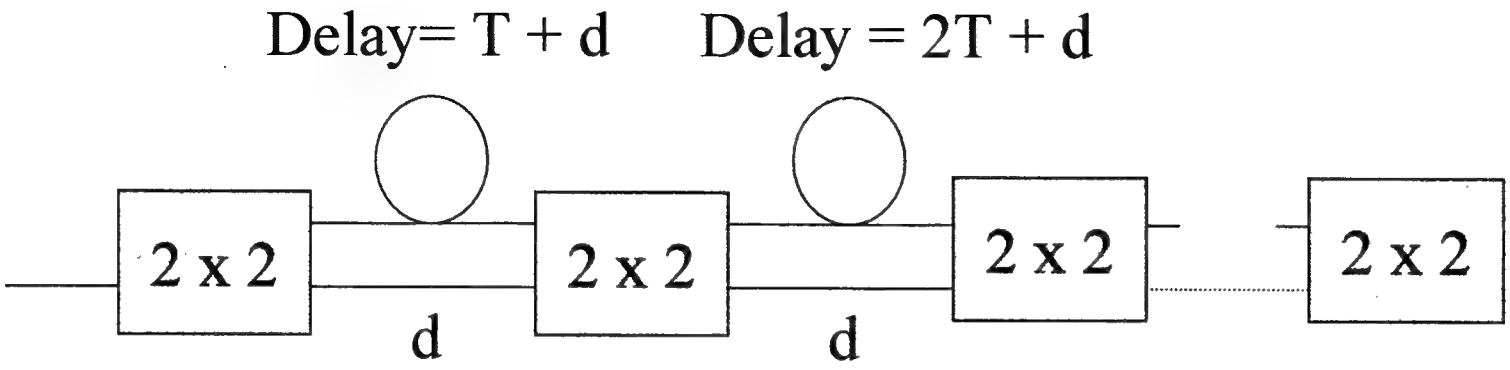
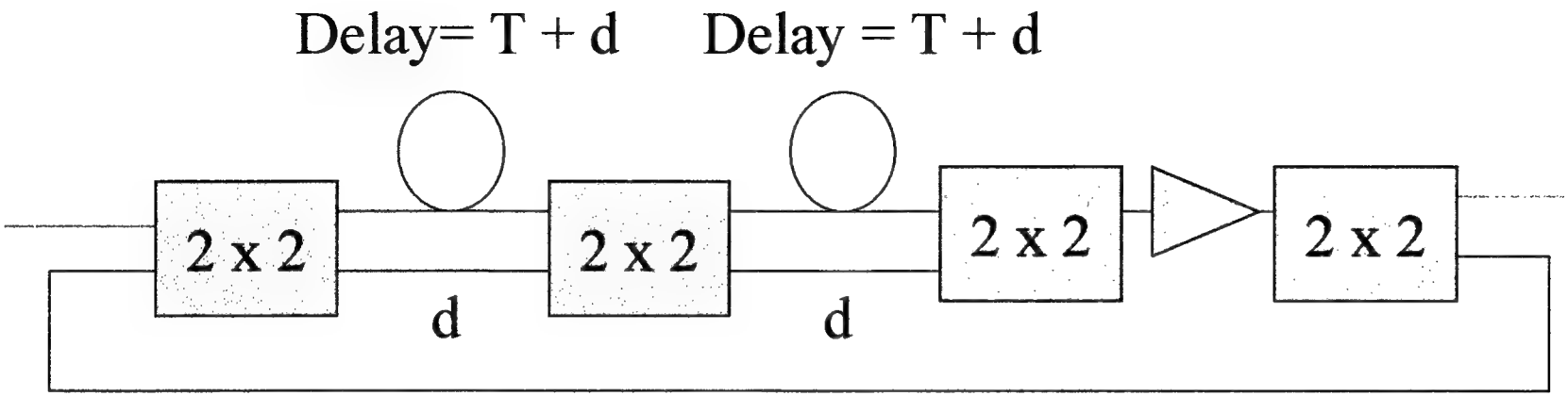


Figure 7a 5 stage buffer: 0-31 T
(predetermined delay)

T= time between adjacent packets
d= minimum delay for each stage



Delay = T- Nd-a

Figure 7b N stage buffer: 0-31 T (non-pre-determined delay)

a= delay through amplifier +
delay through last 2 x 2 crosspoint

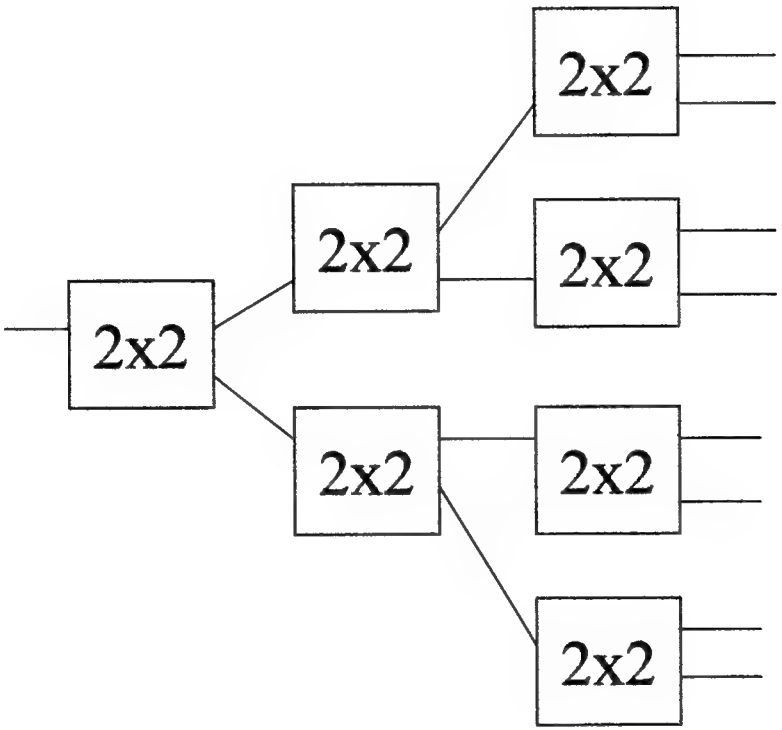


Figure 7c 1 x 8 Banyon Switch

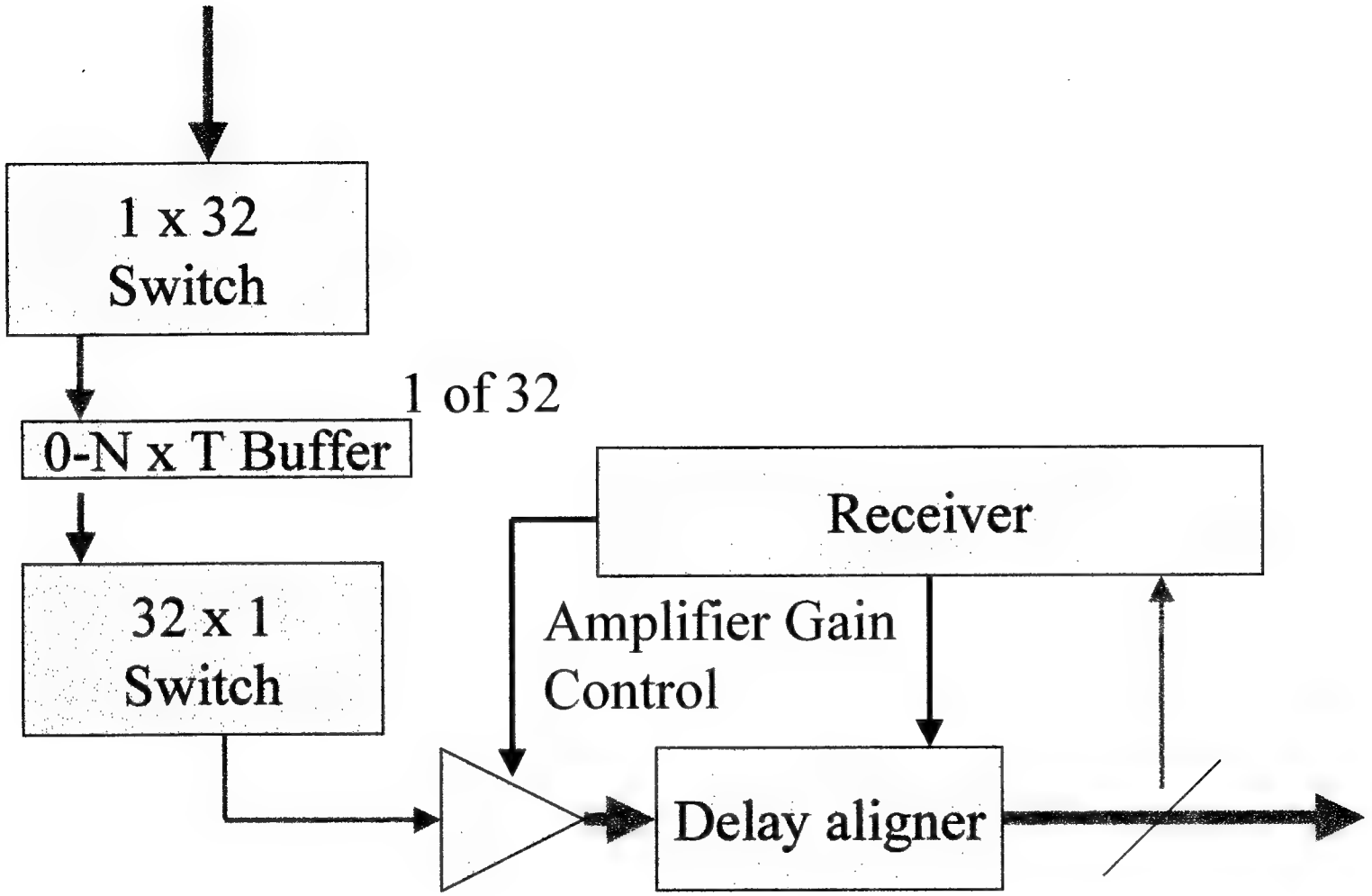


Figure 7d Buffer subsystem

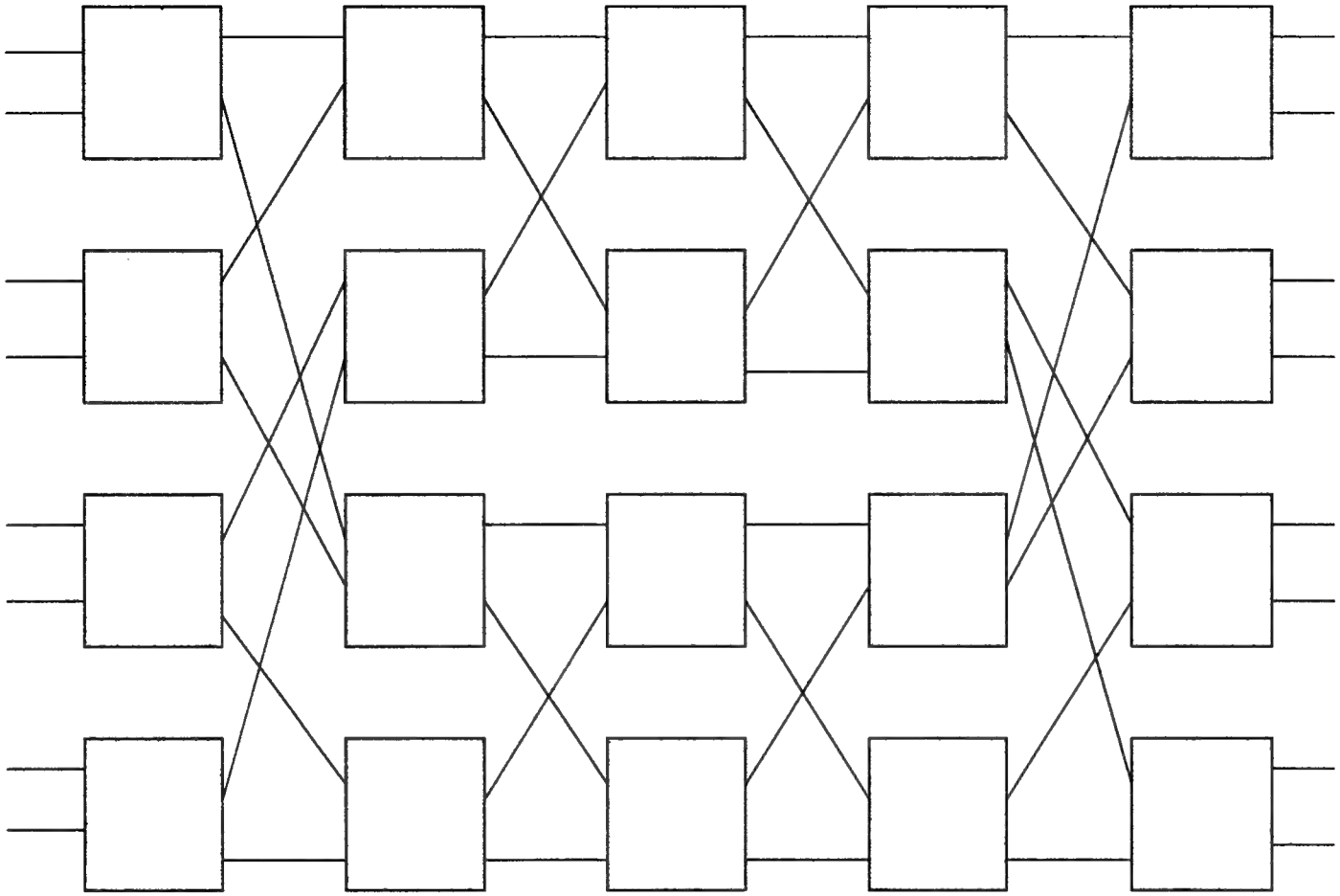


Figure 8: 8 x 8 Benes Matrix

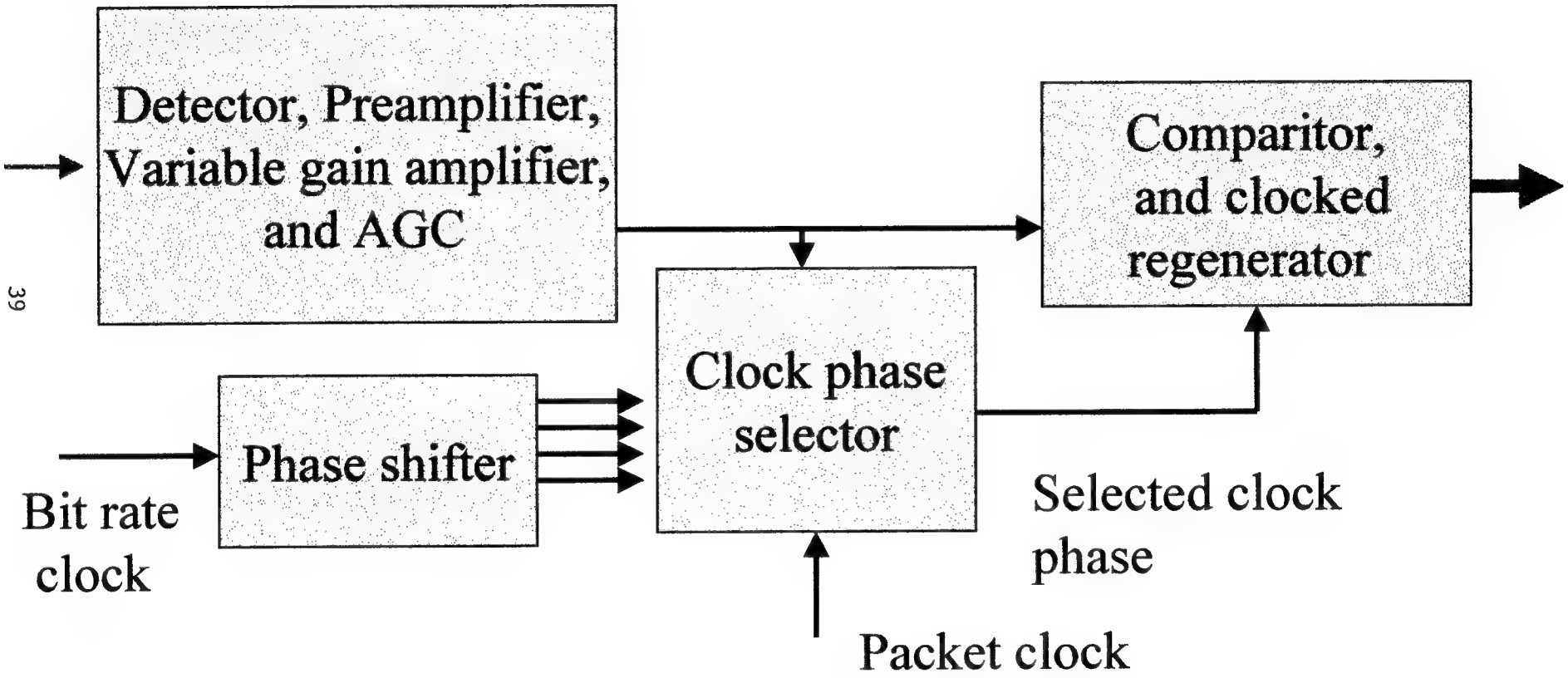


Figure 9 Asynchronous Regenerator

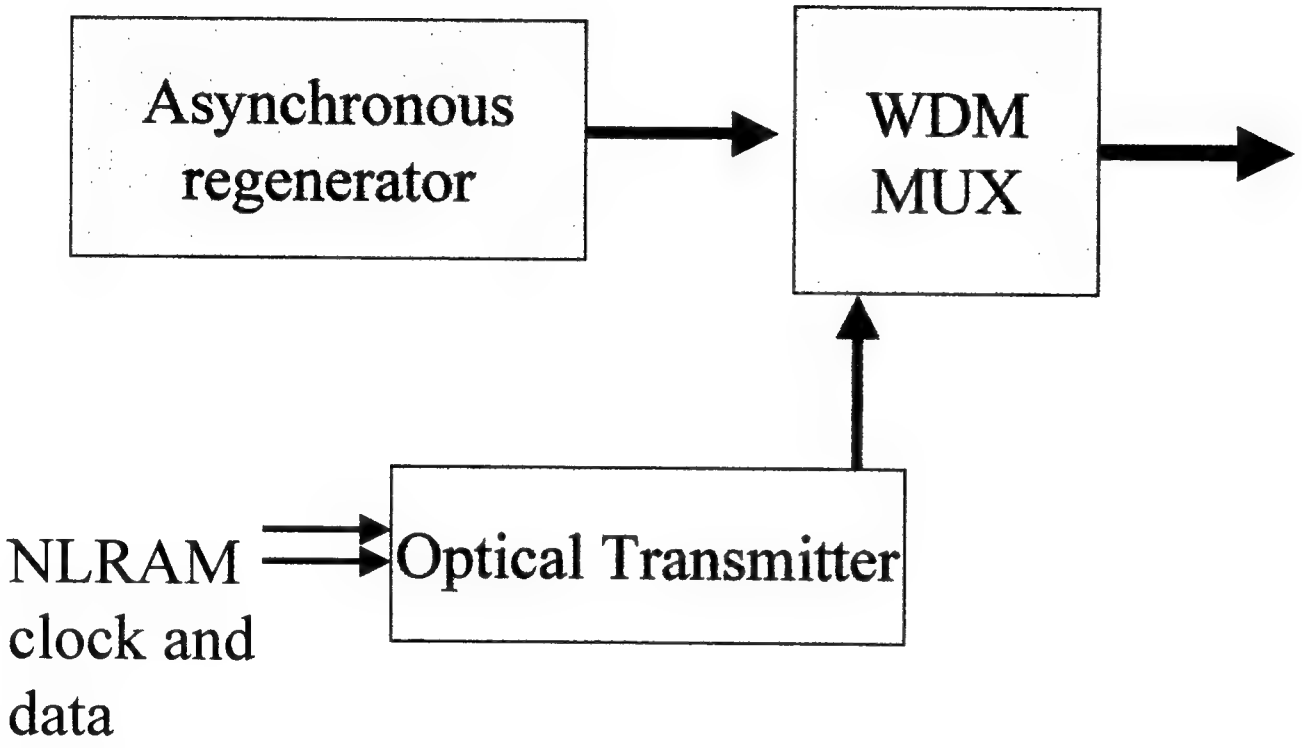


Figure 10a Adding NLRAM: WDM

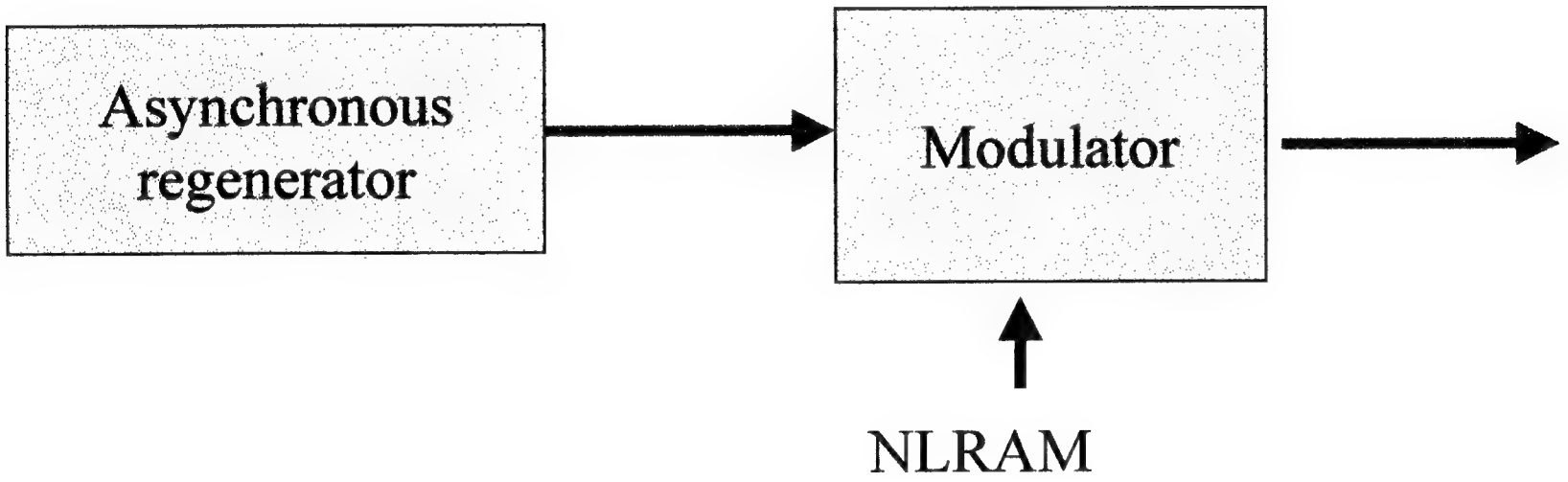


Figure 10b Adding NLRAM: Overlay modulation

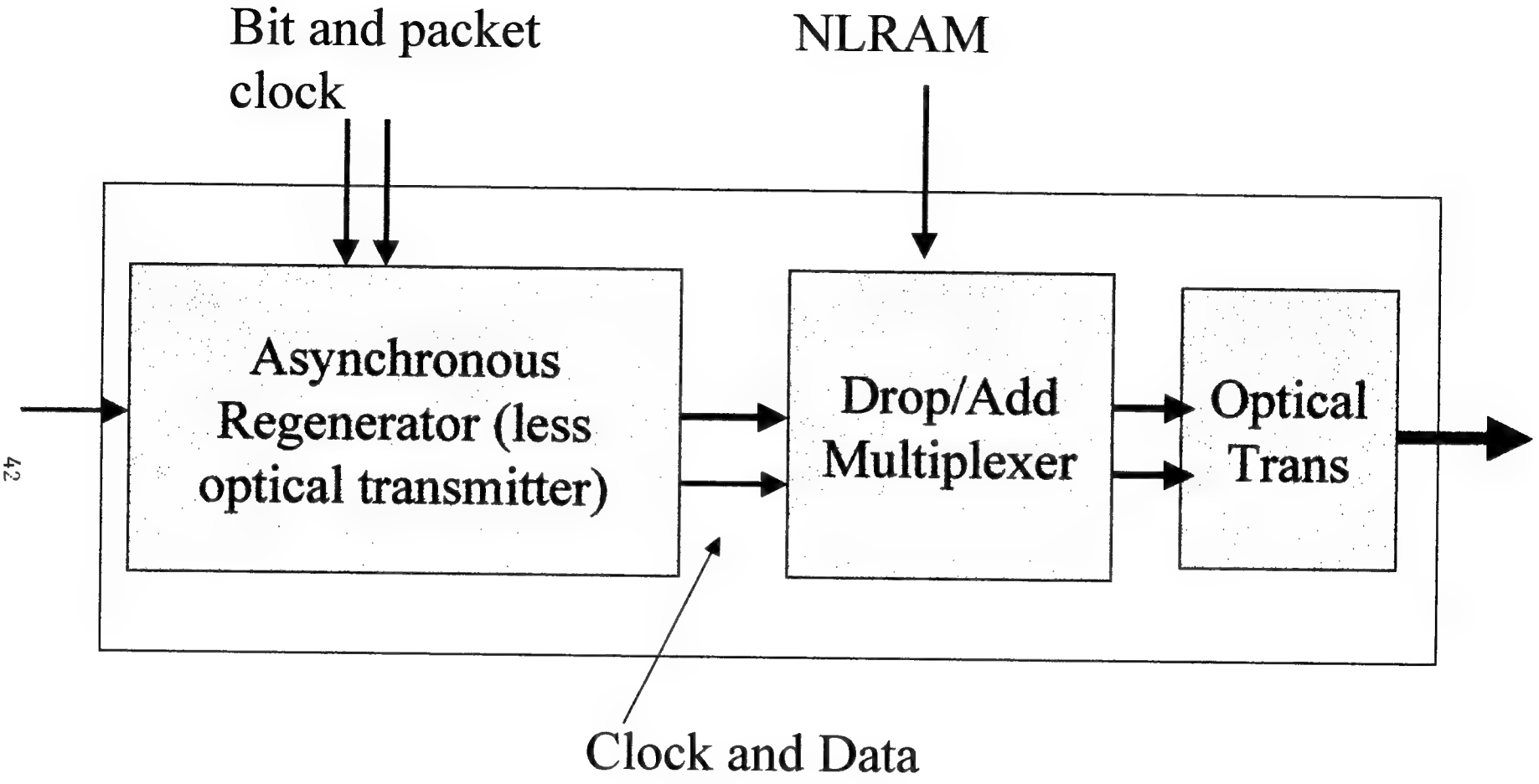
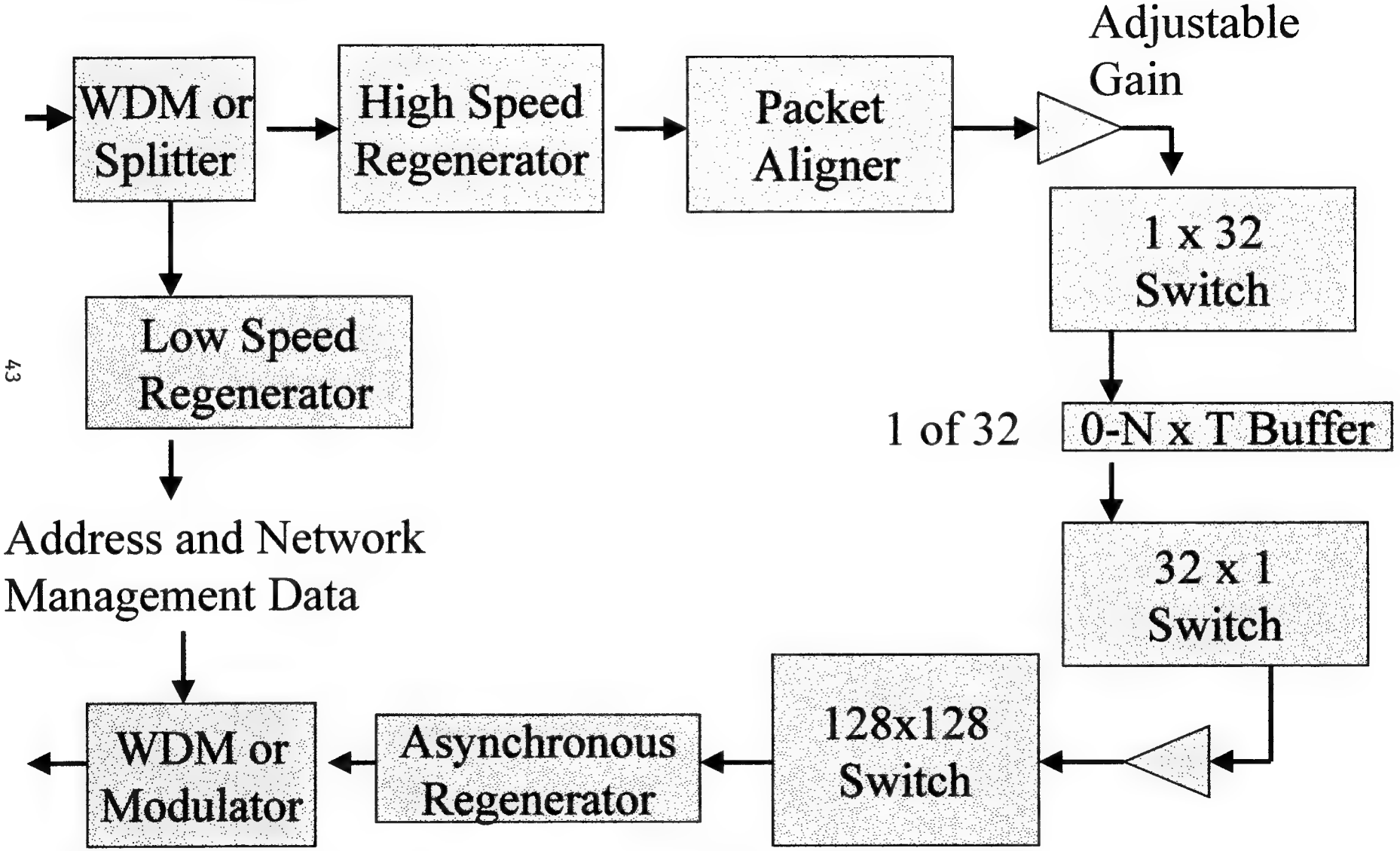


Figure 10c Adding NLRAM: Bit Stream Insertion

Figure 11 Overall Architecture



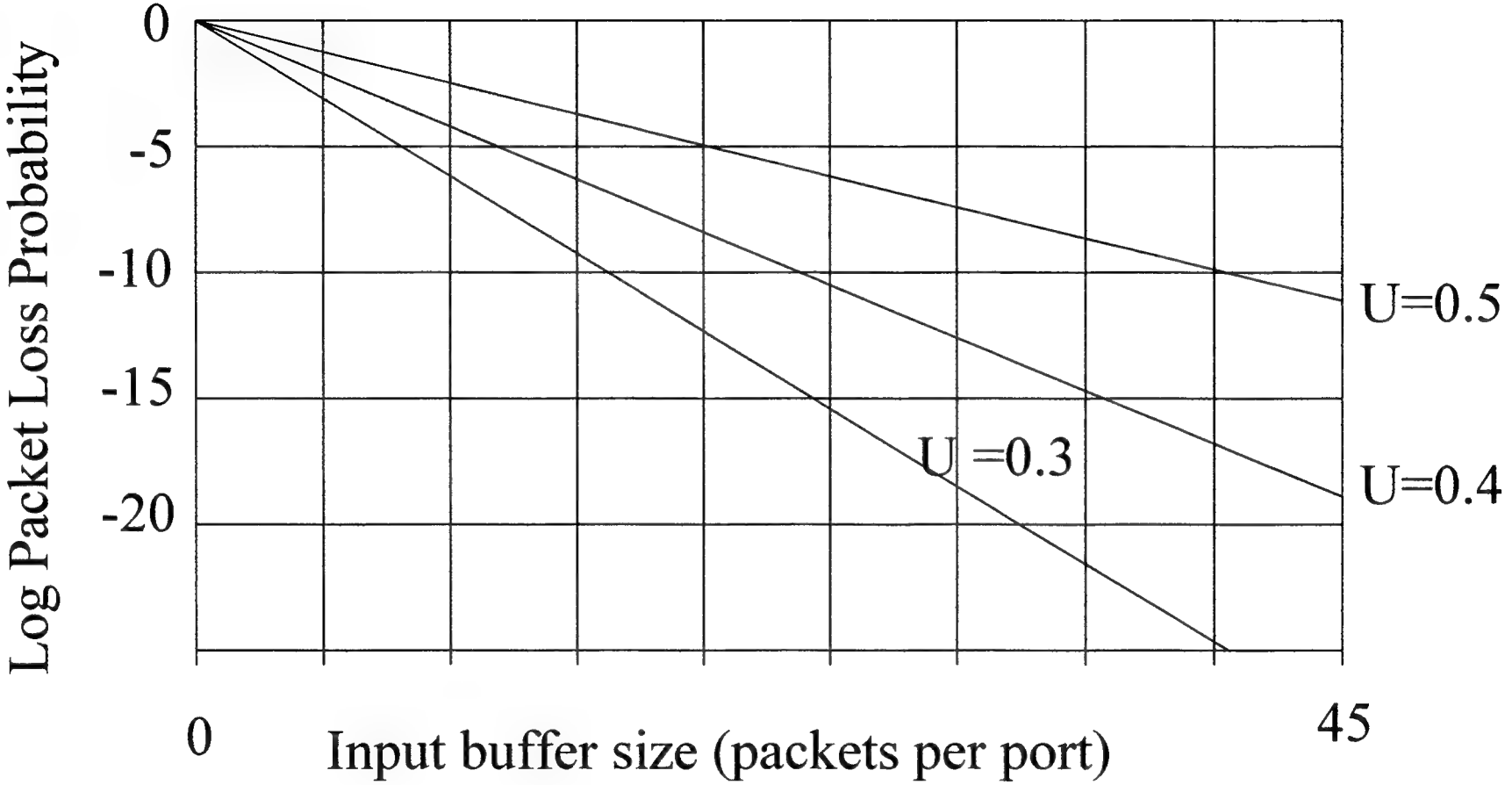


Figure 12: Performance (U= utilization)

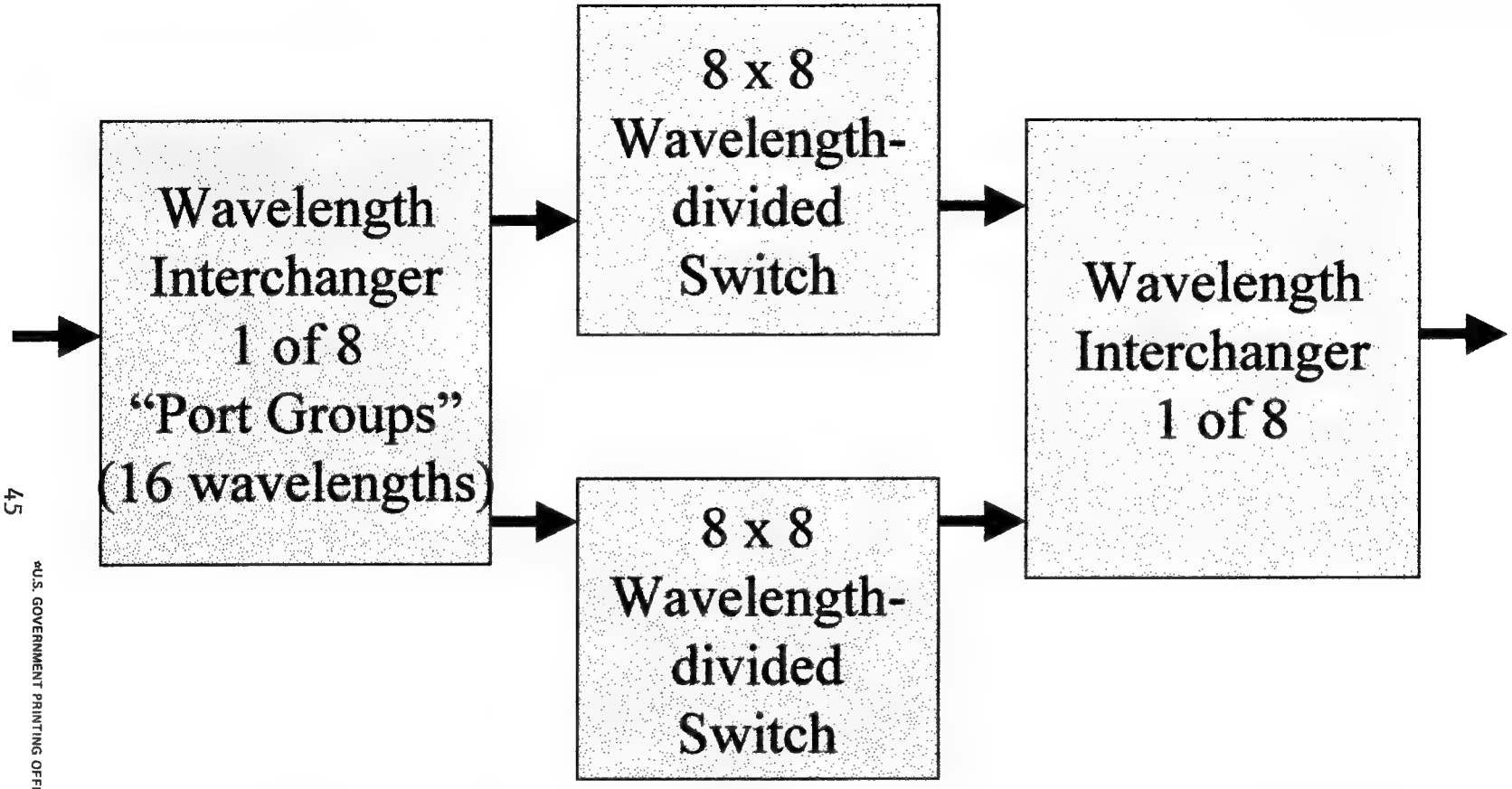


Figure 13 Wavelength-Space-Wavelength Switching Matrix

DISTRIBUTION LIST

addresses	number of copies
AFRL/IFGA ATTN: PRISCILLA CASSIDY 525 BROOKS ROAD ROME, NEW YORK 13441-4505	2
DREXEL UNIVERSITY DEPT OF ELEC & COMPUTER ENGINEERING 3141 CHESTNUT STREET PHILADELPHIS, PA 19104	5
AFRL/IFOIL TECHNICAL LIBRARY 26 ELECTRONIC PKY ROME NY 13441-4514	1
ATTENTION: DTIC-OCC DEFENSE TECHNICAL INFO CENTER 8725 JOHN J. KINGMAN ROAD, STE 0944 FT. BELVOIR, VA 22060-6218	1
DEFENSE ADVANCED RESEARCH PROJECTS AGENCY 3701 NORTH FAIRFAX DRIVE ARLINGTON VA 22203-1714	1
ATTN: NAN PFRIMMER IIT RESEARCH INSTITUTE 201 MILL ST. ROME, NY 13440	1
AFIT ACADEMIC LIBRARY AFIT/LDR, 2950 P.STREET AREA B, BLDG 642 WRIGHT-PATTERSON AFB OH 45433-7765	1
AFRL/MLME 2977 P STREET, STE 6 WRIGHT-PATTERSON AFB OH 45433-7739	1

AFRL/HESC-TDC 1
2698 G STREET, BLDG 190
WRIGHT-PATTERSON AFB OH 45433-7604

ATTN: SMDC IM PL 1
US ARMY SPACE & MISSILE DEF CMD
P.O. BOX 1500
HUNTSVILLE AL 35807-3801

TECHNICAL LIBRARY 00274(PL-TS) 1
SPAWARSYSCEN
53560 HULL ST.
SAN DIEGO CA 92152-5001

CDR, US ARMY AVIATION & MISSILE CMD ?
REDSTONE SCIENTIFIC INFORMATION CTR
ATTN: AMSAM-RD-OR-R, (DOCUMENTS)
REDSTONE ARSENAL AL 35898-5000

REPORT LIBRARY 1
MS P364
LOS ALAMOS NATIONAL LABORATORY
LOS ALAMOS NM 87545

ATTN: D'BORAH HART 1
AVIATION BRANCH SVC 122.10
FOB10A, RM 931
800 INDEPENDENCE AVE, SW
WASHINGTON DC 20591

AFIWC/MSY 1
102 HALL BLVD, STE 315
SAN ANTONIO TX 78243-7016

ATTN: KAROLA M. YOURISON 1
SOFTWARE ENGINEERING INSTITUTE
4500 FIFTH AVENUE
PITTSBURGH PA 15213

USAF/AIR FORCE RESEARCH LABORATORY 1
AFRL/VSOSA(LIBRARY-BLDG 1103)
5 WRIGHT DRIVE
HANSCOM AFB MA 01731-3004

ATTN: EILEEN LADUKE/D460 1
MITRE CORPORATION
202 BURLINGTON RD
BEDFORD MA 01730

OUSD(P)/DTSA/DUTD 1
ATTN: PATRICK G. SULLIVAN, JR.
400 ARMY NAVY DRIVE
SUITE 300
ARLINGTON VA 22202

MISSION
OF
AFRL/INFORMATION DIRECTORATE (IF)

The advancement and application of Information Systems Science and Technology to meet Air Force unique requirements for Information Dominance and its transition to aerospace systems to meet Air Force needs.